Using Machine Learning and Systems-Biology Approaches to Analyse Next-Generation Sequence Data in Cancers

Sutherland, Russel David

Awarding institution:
King's College London

Using Machine Learning and Systems-Biology Approaches to Analyse Next-Generation Sequence Data in Cancers

Russel David Sutherland

A Thesis Submitted for the Degree of Doctor of Philosophy

Department of Medical & Molecular Genetics

King's College London

September 2015

Acknowledgements

I would like to thank my supervisor Dr Richard Dobson for all of the guidance, wisdom, and understanding he has given throughout my PhD. I would like to thank Professor Cathryn Lewis for her encouragement, expertise, and motivational mantras at times of difficulty. Thank you to my previous supervisor Dr Thomas Schlitt, who introduced me to the world of protein-protein interaction networks. Thank you to Dr Anthony Cox, my supervisor at Illumina for all of his support throughout my PhD. Thank you to Dr Adrian Alexa for his advice and for being a good friend during the last four years. I was very fortunate to work with Dr Salvador Diaz-Cano and Dr Jane Moorhead, who always have fantastic ideas.

I gratefully acknowledge the financial support received from the Biotechnology and Biological Sciences Research Council (BBSRC) Cooperative Awards in Science and Engineering (CASE) studentship in collaboration with Illumina (BB/I016287/1).

Thank you to my friends Jack Euesden, Seth Seegobin, Dr Steven Kiddle, and Dr Robert Power for their encouragement and generosity.

I would like to thank my family who have always believed in me. Finally I would like to thank Kirsty Felstead for her relentless positivity, support, and love throughout my PhD.

Statement of work

All of the analyses and programming presented in this thesis have been conducted by Russel Sutherland.

The ideas to identify the mutated genes associated with cancer grade across cancer types (Chapter 3) and to identify the mutated genes that discriminate between five high-order cancer types (Chapter 4.1) came from a collaboration between Russel Sutherland, Dr Richard Dobson, Dr Salvador Diaz-Cano, and Dr Jane Moorhead.

Abstract

The availability of exome sequence data for thousands of cancer samples has enabled the investigation of the sequence-level mutations that contribute to cancer. There is a need for strategies to analyse sequence data to gain new biological and clinical insights. This thesis investigates the use of machine learning and network-based methods to identify the mutated genes associated with important clinical features and cancer types, and to aid candidate gene prioritisation in colorectal cancer, and rheumatoid arthritis.

Firstly, tumour / normal exome sequence data was analysed to identify the mutated genes associated with cancer grade and cancer stage across and within three adenocarcinomas. Tumour grading is an important prognostic indicator which is based upon subjective assessment by pathologists, and is not standardised across cancer types. Despite this, this study found that protein coding mutations within *TP53* were indicative of high grade status across three adenocarcinomas once adjusted for age, gender, stage, and tumour type.

Secondly, Random Forest models were used to identify the mutations that discriminate each of five high-order cancer types. Based on this work a Random Forest approach was used to investigate whether exome sequence data could be used to assign cancers to their tissue of origin without prior knowledge, for future use as a classifier for cancers of unknown primary origin.

Finally, a network-based method to perform candidate disease gene prioritisation called 'k-pseudo cliques analysis' was developed. The method identifies sets of highly interacting proteins that are enriched for low gene-level p-values. In tests, the identified gene sets outperformed a univariate test for general cancer gene enrichment. As part of the final chapter a network-based method called 'Region Growing Analysis' was used to perform candidate disease gene prioritisation of rheumatoid arthritis genome-wide association study data.

The findings and methods developed in this thesis can provide insights to the genetic correlates of cancer phenotypes and suggest new candidate disease genes.

Abbreviations

AGSTP	Age, gender, stage, tumour type, proteins
AIC	Akaike's Information Criterion
AML	Acute myeloid leukaemia
AUC	Area under the curve
BLCA	Bladder urothelial carcinoma
BRCA	Breast adenocarcinoma
CNV	copy number variant
COAD	Colon adenocarcinoma
COADREAD	Colon and rectal adenocarcinoma
COSMIC	Catalogue of Somatic Mutations in Cancer
CUP	Cancers of unknown primary
FDR	False discovery rate
GBM	Glioblastoma multiforme
GO	Gene Ontology
Gr	Grade
GWAS	Genome-wide association study
HNSC	Head and neck squamous cell carcinoma
HuPPI2	Human Protein Protein Interaction Network 2
ICGC	International Cancer Genome Consortium
Indel	Insertion, or deletion mutation
KEGG	The Kyoto Encyclopedia of Genes and Genomes
KIRC	Kidney renal clear cell carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MHC	Major histocompatibility complex
MSI-H	Micro satellite instability-high
MSI-L	Micro satellite instability-low
MSI-S	Micro satellite instability-stable
NGS	Next generation sequencing
nmMDS	non-metric Multi-Dimensional Scaling
OMIM	Online Mendelian Inheritance in Man
OR	Odds ratio
OV	Ovarian serous carcinoma
PC	Principal component
PCA	Principal component analysis
PPI	Protein-protein interaction (network)
RA	Rheumatoid arthritis
READ	Rectal adenocarcinoma
RGA	Region Growing Analysis
ROC	Receiver operator characteristic
SCNV	Somatic copy number variant
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
TCGA	The Cancer Genome Atlas
UCEC	Endometrial carcinoma
WTCCC	Wellcome Trust Case-Control Consortium

Presentations and Posters

Throughout the course of the research undertaken as part of the PhD the following presentations and posters were presented at academic conferences.

Presentations

2012 **Region Growing Analysis: A de novo pathway discovery tool.**
European Conference on Computational Biology Student Council Symposium (ECCBSCS) Student Council Symposium, Basel, Switzerland.

2014 **Using exome sequence data and Random Forest analysis to identify functional mutation signatures of 5 cancer differentiation subtypes.**
ISMB Student Council Symposium, Boston, USA.

Predicting adenocarcinoma tumour grade using exome sequence data and clinical data.
International Society for Computational Biology Regional Student Group (ISCBRSG) Meeting, Treforrest, UK.

Posters

2012 **Region Growing Analysis: A de novo pathway discovery tool.**
European Conference on Computational Biology (ECCB). Basel, Switzerland

Prioritising mutated genes using protein interaction networks.
Cancer Research Institute Symposium, Cambridge, UK.

2014

Using exome sequence data and Random Forest analysis to identify functional mutation signatures of 5 cancer differentiation subtypes.

Next Generation Sequencing Conference, Barcelona, Spain.

Using exome sequence data and Random Forest analysis to identify functional mutation signatures of 5 cancer differentiation subtypes.

Intelligent Systems for Molecular Biology (ISMB), Boston, USA.

Predicting adenocarcinoma tumour grade using exome sequence data and clinical data.

ECCB, Strasbourg, France.

Contents

1 Introduction **17**

1.1 Motivation . . . 18

1.2 An overview of cancer biology . . . 19

1.3 Discovering genes that contribute to cancer using exome sequence data. . . . 21

1.3.1 Characteristic cancer mutations . . . 21

1.3.2 Heterogeneity in cancer . . . 22

1.4 The Cancer Genome Atlas . . . 24

1.5 The Pan-Cancer analysis project . . . 24

1.6 Cancer staging and tumour grading . . . 25

1.7 High-order cancer types in the Pan-Cancer dataset . . . 26

1.7.1 Cancers of unknown primary origin . . . 27

1.8 An outline of exome sequencing . . . 28

1.8.1 Types of mutations uncovered by exome sequencing . . . 29

1.8.2 Exome-sequencing in cancer studies . . . 30

1.9 Clinical practice and next-generation sequencing . . . 31

1.10 An overview of machine learning . . . 33

1.10.1 Supervised and unsupervised methods . . . 33

1.10.2 Evaluation of machine learning methods . . . 33

1.10.3 Decision tree classifiers . . . 38

1.10.4 Random Forest . . . 39

1.11 How protein interaction networks can be used to identify new disease gene candidates. . . . 40

1.11.1 Protein-protein interaction networks . . . 40

1.11.2 An overview of graph theory . . . 41

1.11.3 Network properties of disease genes . . . 45

1.11.4 Predicting disease genes using networks . . . 46

1.12 Study aims . . . 48

1.12.1 Specific aims of the thesis . . . 49

2 Exploratory analysis of The Cancer Genome Atlas and Pan-Cancer colorectal cancer datasets 51
2.1 Introduction . . . 52
2.1.1 Confounding factors . . . 52
2.1.2 Micro-satellite instability . . . 54
2.2 Methods . . . 55
2.2.1 Binary mutation matrix . . . 57
2.2.2 Non-Metric Multi Dimensional Scaling . . . 57
2.2.3 Identifying features correlated with the first principal component . . . 58
2.2.4 Over-representation analyses to discover potential study design problems . . . 58
2.3 Results . . . 59
2.3.1 non-metric Multi-Dimensional Scaling . . . 59
2.3.2 Metadata correlations with the first principal component. . . . 62
2.3.3 Class imbalance analysis . . . 63
2.4 Discussion . . . 63
2.5 Conclusion . . . 64

3 Predicting cancer grade and stage across three types of adenocarcinoma using exome sequence data 65
3.1 Introduction . . . 66
3.2 Methods . . . 68
3.2.1 The data set . . . 69
3.2.2 Data pre-processing . . . 69
3.2.3 Mutated protein and variant frequency feature selection . . . 70
3.2.4 Final model building . . . 71
3.3 Results . . . 73
3.3.1 Cross-adenocarcinoma grade classification . . . 73
3.3.2 Stage classification . . . 79
3.4 Discussion . . . 82
3.5 Conclusion . . . 85
3.6 Supplementary materials . . . 86

4 Predicting cancer types using TCGA exome sequence data and Random Forest analysis. 99

4.1 Using exome sequence data and Random Forest analysis to identify the functional mutation patterns of five high-order cancer types 101

4.1.1 Introduction . . . 102

4.1.2 Methods . . . 105

4.1.2.1 The data set . . . 105

4.1.2.2 Definition of the five cancer high-order cancer types . . . 105

4.1.2.3 Creating the binary mutation matrix . . . 106

4.1.2.4 Training set and test set definition. . . . 106

4.1.2.5 Random Forest prediction models . . . 107

4.1.3 Results . . . 113

4.1.3.1 Univariate analyses . . . 113

4.1.3.2 Random Forest prediction models . . . 113

4.1.3.3 KEGG disease pathway enrichment . . . 118

4.1.4 Discussion . . . 121

4.1.5 Conclusions . . . 124

4.1.6 Supplementary Materials . . . 125

4.2 Working towards predicting the origin of cancers of unknown primary using whole exome sequence data. 138

4.2.1 Introduction . . . 139

4.2.2 Methods . . . 141

4.2.2.1 The Pan-Cancer dataset . . . 141

4.2.2.2 Training set sampling . . . 141

4.2.2.3 Random Forest model building . . . 143

4.2.2.4 Classifier performance measures . . . 145

4.2.3 Results . . . 146

4.2.3.1 The down-sampled classifier . . . 146

4.2.3.2 The up-sampled classifier . . . 146

4.2.4 Discussion . . . 149

4.2.5 Conclusion . . . 150

5 Candidate disease gene prioritisation for complex diseases using network-based methods 151

5.1 Network-based disease gene prioritisation using colorectal cancer exome sequence data 153

5.1.1 Introduction . . . 154

5.1.2 Methods . . . 157

5.1.2.1 Mutation and network data . . . 157

5.1.2.2 MutSigCV analysis . . . 157

5.1.2.3 The HuPPI2 network . . . 157

5.1.2.4 Enumeration of pseudo cliques . . . 158

5.1.2.5 K-pseudo clique definition . . . 158

5.1.2.6 Network permutation . . . 159

5.1.2.7 KEGG enrichment analysis . . . 162

5.1.3 Results . . . 164

5.1.3.1 Network statistics . . . 164

5.1.3.2 Network permutation results . . . 164

5.1.3.3 Number and type of pseudo cliques identified at each density threshold . . . 165

5.1.3.4 Median univariate test statistic permutation tests . . . 166

5.1.3.5 KEGG enrichment analysis . . . 166

5.1.4 Discussion . . . 173

5.1.4.1 Additional colorectal cancer associated genes in the α=0.95 and k=3 k-pseudo cliques . . . 173

5.1.4.2 Cancer related signalling pathways . . . 174

5.1.4.3 Utility for complex disease genetics discovery . . . 175

5.1.4.4 K-pseudo cliques analysis future improvements . . . 175

5.1.5 Conclusion . . . 177

5.2 Prioritising rheumatoid arthritis candidate disease genes using Region Growing Analysis. 179

5.2.1 Introduction . . . 179

5.2.1.1 Network-based analysis of GWAS data . . . 181

5.2.1.2 Study outline . . . 182

5.2.2 Methods . . . 183

5.2.2.1 The WTCCC RA dataset . . . 183

5.2.2.2 Gene score generation . . . 183

5.2.2.3 Region Growing Algorithm . . . 184

5.2.3 Results . . . 186
5.2.4 Discussion . . . 190
5.2.5 Conclusions . . . 191
5.2.6 Supplementary Materials . . . 193

6 Conclusions 195

6.1 Overview of the thesis . . . 196
6.2 Research aims revisited . . . 198
6.3 Contributions . . . 201
6.4 Limitations . . . 201
6.5 Future work . . . 202
6.6 Concluding remarks . . . 203

Bibliography 205

List of Figures

1.8.1 Sequencing by synthesis . . . 29
1.10.1 k-fold cross validation procedure . . . 35
1.10.2 ROC curve example . . . 38
1.11.1 Protein interaction network graph . . . 42
1.11.2 Vertex measures . . . 44
2.2.1 Exploratory analysis pipeline . . . 56
2.3.1 TCGA December 2012 colorectal nmMDS plots . . . 61
2.3.2 TCGA June 2012 colorectal nmMDS plots . . . 61
2.3.3 Pan-Cancer colorectal nmMDS plots . . . 62
3.2.1 Grade and stage classification analysis pipeline . . . 72
3.3.1 Across cancer grade classification ROC curves. . . . 75
3.3.2 Percentage of low-grade and high-grade samples carrying protein coding mutations in each *TP53* exon. . . . 77
3.3.3 Endometrial carcinoma grade classification ROC curves . . . 79
3.3.4 Endometrial carcinoma sample mutation frequency. . . . 79
3.3.5 Stage classification across cancers ROC curves . . . 80
3.6.1 Variant types are non-normally distributed across cancer grade . . . 86
3.6.2 Endometrial carcinoma stage classification ROC . . . 96
3.6.3 Ovarian carcinoma ROC curves . . . 96
3.6.4 Renal carcinoma stage classification ROC curves . . . 97

4.1.2.1 Random Forest with recursive feature elimination and 5 x 10 fold cross validation. . . . 110
4.1.3.1 Random Forest classification model ROC curves . . . 117
4.1.6.1 Pairwise Random Forest classification accuracy plots . . . 126
4.1.6.2 Adenocarcinoma / squamous cell carcinoma heat map . . . 128
4.1.6.3 Adenocarcinoma / urothelial carcinoma heat map . . . 129
4.1.6.4 Squamous cell carcinoma / urothelial carcinoma heat map . . . 130
4.1.6.5 Adenocarcinoma / glioblastoma heat map . . . 131

4.1.6.6 Squamous cell carcinoma / glioblastoma heat map . . . 132
4.1.6.7 Urothelial / glioblastoma heat map . . . 133
4.1.6.8 Adenocarcinoma / leukemia heat map . . . 134
4.1.6.9 Squamous cell carcinoma / leukemia heat map . . . 135
4.1.6.10 Urothelial / leukemia heat map . . . 136
4.1.6.11 Glioblastoma / leukemia heat map . . . 137

4.2.2.1 Cancer of unknown origin classifier building pipeline . . . 144
4.2.3.1 Cancer of unknown origin classification ROC curves . . . 148

5.1.2.1 The k-pseudo clique search process Three pseudo cliques that overlap by 3 vertices are combined in to a single k-pseudo clique. . . . 159
5.1.3.1 HuPPI2 network degree distribution . . . 164
5.1.3.2 Distributions of the absolute degree difference of permuted vertex labels and original vertex labels using two permutation methods. . . . 165
5.1.3.3 Pseudo-clique size frequencies from change across density thresholds α=0.95 to α=0.75. 166
5.1.3.4 Histograms of significant k-pseudo cliques . . . 167
5.1.3.5 HuPPI α=0.95, 0.90, and 0.85 k=3 significantly mutated subnetworks. . . . 172

5.2.2.1 Visualising the steps in the RGA algorithm and its application to a protein interaction network. . . . 186
5.2.3.1 Max T permutation results and significant regions . . . 187
5.2.3.2 Top Q permutation results and significant region . . . 188
5.2.3.3 Mean T permutation results and significant region . . . 189

List of Tables

1.10.1 Binary classification statistics . . . 37

2.1.1 Sequencing technology and analysis pipelines in TCGA and Pan-Cancer datasets . . . 53

2.3.1 First principal component correlations with metadata in each data set. . . . 62

3.2.1 Grade and stage classification model abbreviations . . . 68

3.2.2 Number of proteins retained after frequency filter . . . 71

3.3.1 Across cancer grade classification statistics . . . 76

3.3.2 Across cancer stage classification statistics . . . 81

3.6.1 Grade and stage demographics table . . . 87

3.6.2 Descriptive statistics and class imbalance tests across and within cancers . . . 88

3.6.3 Stage assignment frequencies stratified by grade . . . 89

3.6.4 Grade assignment stratified by tumour stage (low/high) . . . 89

3.6.5 Complete across cancer grade classification models . . . 90

3.6.6 Complete across cancer stage classification models . . . 91

3.6.7 Cancer grade prediction models within cancers . . . 92

3.6.8 Cancer stage prediction models within cancers . . . 93

3.6.9 Endometrial carcinoma grade and stage classification statistics . . . 94

3.6.10 Renal cell carcinoma grade and stage classification statistics . . . 95

3.6.11 Ovarian carcinoma grade and stage classification statistics . . . 98

4.1.2.1 High-order cancer type assignment . . . 106

4.1.2.2 Number of proteins used for model building, and training and test set sizes. . . . 107

4.1.3.1 High-order cancer type demographics . . . 113

4.1.3.2 Median frequencies of mutations, and univariate tests across cancer types . . . 114

4.1.3.3 Random Forest classification model statistics . . . 116

4.1.3.4 Random Forest model features and variable importance measures . . . 119

4.1.6.1 cancer stage distribution across cancers . . . 125

4.1.6.2 cancer stage distribution across high-order cancer types . . . 125

4.2.2.1 Two third training set sizes for each cancer type used to build the CUP classifiers . . . 143

4.2.3.1 Test set performance statistics for the down-sampled and up-sampled CUP classifiers 147
4.2.3.2 Random Forest classifier feature importance 149

5.1.3.1 K-pseudo cliques with a lower than expected median MutSigCV p-value 170
5.1.3.2 KEGG pathway enrichment validation of MutSigCV and k-pseudo cliques analysis . . 171
5.1.5.1 KEGG enrichment of significantly mutated k-pseudo cliques larger than ten proteins. 178

5.2.6.1 Genes within the significant 27 gene region found using a rank threshold of α=350 based on maxT gene-wide empirical p-value ranks. 193
5.2.6.2 Genes within the significant region of 10 genes found using a rank threshold of α=200 and topQ gene-wide empirical p-value ranks. 194
5.2.6.3 Genes within the significant 6 gene region found using a rank threshold of α=60 based on meanT gene-wide empirical p-value ranks 194

Chapter 1

Introduction

1.1 Motivation

In the last decade there has been an explosion in the amount of DNA sequence data available. This is due to the development of next generation sequencing technologies that can generate gigabases of sequence data and provide sequence level characterisation of an individual's genome. The decline in the cost of this technology has allowed the successful large scale study of sequence level variation of cancers. However, the complete set of mutated genes that are associated with cancer types and important clinical features has not yet been discovered. Throughout this thesis I will investigate these issues using machine learning and systems-biology methods to analyse mutations discovered from whole exome sequence analysis of thousands of tumour samples.

I will use logistic regression analyses in Chapter 3 to identify the mutated genes associated with tumour grade and cancer stage across three types of cancer. Tumour grading is a subjective measure of prognosis and is not standardised across cancer types. By establishing mutated genes associated with tumour grade across cancer types the foundations for a single tumour grading system can be established.

The Pan-Cancer analysis of twelve cancers is composed of five high-order cancer types (adenocarcinomas, squamous cell carcinomas, urothelial carcinomas, blood cancers, and cancers of the central nervous system). However, the sets of mutated genes that discriminate between these high-order types have not been identified. I will use Random Forest analyses in Chapter 4.1 to identify sets of mutated genes that discriminate each high-order cancer type from others, and to uncover new genes that are important for the progression for each cancer type.

A further Random Forest model will be applied to address the problem of cancers of unknown primary origin in Chapter 4.2. Around three percent of cancers are diagnosed as stage IV with an unknown primary origin. To provide a prediction of the origin of the primary tumour may aid patient care. I will use multi-class Random Forest models to assign the Pan-Cancer cancer types to their tissue of origin using exome sequence data.

The problem of prioritising genes that contribute to cancer and other complex diseases

is addressed in Chapter 5. Tests to identify genes contributing to cancer conducted at the gene level may be underpowered to detect some of the contributing genes due to genetic heterogeneity, where samples may carry mutations in different genes that have the same phenotypic effect. In Chapter 5.1 I will describe how I developed a method called *k-pseudo cliques analysis* which can prioritise disease genes using gene level test statistics and a protein interaction network. A further demonstration of the utility of protein interaction networks for prioritising candidate disease genes will be provided by applying *Region Growing Analysis* to a rheumatoid arthritis genome-wide association study dataset in Chapter 5.2.

Finally I will conclude with the findings from these investigations. I will discuss the limitations of these analyses and how they can be further refined.

Most of the analyses in this thesis relate to cancers. First I will give some background to cancer biology. I will go on to describe provide an overview of areas important for each chapter in this thesis including cancer grading and staging, high order cancer types, and how protein protein interaction networks and high throughput genomic data can be used to prioritise disease genes. The introduction will conclude with an outline of the aims of the thesis.

1.2 An overview of cancer biology

Cancer is characterised by uncontrolled cell growth and proliferation, invasion in to surrounding tissue, and the eventual spread to multiple sites across the body. When human cells die, or become damaged, other cells grow and divide to replace them. Cancer may occur if this process is disrupted. Cancers are typically treated using surgery, chemotherapy, radiotherapy, targeted therapy, immunotherapy, or hormone therapy. Chemotherapy involves using drugs, often in combination, to kill cancer cells. Radiotherapy uses x-rays and other high-energy radiation to kill cancer cells using a "beam" of radiation, or by introducing radioactive material in to the body. Targeted therapy, using drugs, or monoclonal antibodies, that specifically target cancer cells can have fewer side-effects than other treatments. Hormone therapy can be given to stop, or slow the growth of hormone sensitive cancers before surgery.

Cancers are caused by genetic mutation, but many factors influence its occurrence. Genetic, environmental, and epigenetic factors are all known to be important across the many types of cancer. The well known environmental factors include smoking, which is associated with a 10-fold increase in lung cancer mutation frequency in comparison to non-smokers (Govindan et al., 2012). Epigenetic alterations, such as DNA methylation, are known to be important for the progression of certain types of colorectal cancer where more than 10 percent of genes can be differentially methylated (Beggs et al., 2013).

For over 100 years cancer has been thought to be a disease of the genome (Boveri, 2008), but only in the 1970s were researchers successful in showing this to be true (Stehelin et al., 1976). There are heritable genetic factors that contribute to cancer (germ-line mutations) and confer an increased risk of developing cancer at some point during a person's lifetime (Pomerantz & Freedman, 2011). Indeed, there is an increased rate of cancer among close relatives of people with cancer (Pomerantz & Freedman, 2011).

All cancers arise due to somatic mutations that arise in the DNA of cancer cells. However, not all somatic mutations will have contributed to the development of the cancer. The mutations which have played some part in the development of the cancer are termed *driver mutations* and there are corresponding *driver genes* in which the driver mutations most commonly occur (Stratton et al., 2009). The somatic mutations that have played no part in the development of the cancer are known as *passenger mutations*. The discrimination between driver and passenger mutations is at the heart of cancer genomics.

The identification of sequence-level variation in cancers was only possible due to the completion of the human genome project (Lander et al., 2001; Venter et al., 2001). The draft human reference sequence was instrumental to our understanding of cancer. However, the development of next-generation sequencing (NGS) technologies, that could generate gigabases of sequence information in a very short time, allowed cancer researchers to characterise sequence-level variation in cancers at a scale that was not possible using classical Sanger sequencing techniques (Sanger & Coulson, 1975). As the cost of next-generation sequencing fell in the mid-2000s, the field of cancer genomics appeared (Mardis, 2011).

Cancer genomics involves the analysis of large datasets, of possibly millions of biomolecular features and thousands of samples in an hypothesis free manner. Its introduction presented a drastic change to the exclusively hypothesis driven approach used in cancer research up to that point. The most high profile whole-'omics approach in cancer research to date has been The Cancer Genome Atlas (TCGA) project, whose aim it was, across 27 research centres, to "accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing".

The Cancer Genome Atlas has so-far collected data on 34 different cancer types, across six different genomic technologies including: full exome sequence, single nucleotide polymorphism (SNP), methylation data, messenger RNA sequencing (mRNA-seq), messenger RNA (mRNA) expression, and micro-RNA (miRNA) sequencing (Weinstein et al., 2013). In addition to the genomic profiling, clinical features were also collected from samples, including cancer stage, cancer grade, age at onset, gender, five-year survival data, and other cancer-type specific clinical features.

The success of TCGA has lead to the establishment of the Genomics England project to conduct whole-genome sequencing on 100 000 genomes, 50 000 of which will be focussed on cancer. The main aim of Genomics England is to benefit patients of rare genetic disorders and cancers, to bring about new medical insights, and to establish robust infrastructure to include NGS technology as part of routine clinical testing in the National Health Service (NHS) (Caulfield et al., 2015).

1.3 Discovering genes that contribute to cancer using exome sequence data.

1.3.1 Characteristic cancer mutations

Single nucleotide variants (SNVs) are the most common type of mutation discovered in cancer exome sequencing. However, mutations much larger than SNVs occur in cancer.

There are somatic copy number variants (SCNVs), where sections of DNA are repeated, or truncated. In colorectal cancer, the amplification of some SCNVs known as micro-satellites is associated with prognosis and is indicative of a malfunction in the DNA mismatch-repair pathway (Muzny et al., 2012). Translocations also occur, where sections of chromosomes are re-ordered, or attached to other chromosomes. In some cases chromosomes are shattered and DNA strand breaks are repaired among random segments of chromosomes in a process known as chromothripsis (Stephens et al., 2011). Regions of local hypermutation where the mean inter-mutation distance between six or more mutations is less than 1000 base-pairs describe a phenomenon called Kataegis (Alexandrov et al., 2013; Nik-Zainal et al., 2012).

1.3.2 Heterogeneity in cancer

When searching for genes that contribute to cancer using NGS data, heterogeneity within samples and across samples can cause some problems which are outlined below.

1.3.2.1 Inter-patient heterogeneity

Two patients may have cancer in the same tissue, but the aetiology and molecular causes of their cancers may be completely different. In most of The Cancer Genome Atlas comprehensive molecular analyses to date very few genes are mutated across more than 10 percent of samples in each cancer type (Vogelstein et al., 2013). Many of the genes mutated at low frequencies in one cancer are often mutated at higher frequencies in another cancer (Vogelstein et al., 2013). Genes that carry mutations in more than one sample, but still relatively few samples are the norm in the large scale cancer studies (Wood et al., 2007; Kandoth et al., 2013a).

The Cancer Genome Atlas studies have used methods to identify significantly mutated genes where the mutation frequency in particular genes is increased above a background mutation rate across a group of samples (Lawrence et al., 2013; Dees et al., 2012). These methods may be underpowered to detect mutated genes that contribute to cancer due to inter-patient genetic heterogeneity (Leiserson et al., 2013b; Vogelstein et al., 2013), where

each patient has slightly different molecular alterations that lead to the same cancer outcome. There may be sufficient power to detect the genes frequently mutated across cancer subtypes, but not the causative genes that are specific to subtypes.

One way to address this problem may be to consider that cancer could be driven by a mutation to any gene within a molecular pathway, any disruption to which causes a cancer phenotype. Wood et al. (2007) suggested that the large number of infrequently mutated genes in breast cancer, may in fact reflect a smaller number of cell-signalling pathways. A pathway based approach may be key in identifying the genes that contribute to cancer progression. Such an approach is explored in Chapter 5.1 on page 153.

1.3.2.2 Intra-tumour heterogeneity

The most popular methods to discover cancer driver genes in large scale sequencing studies use data derived from tumour biopsies (Dees et al., 2012; Lawrence et al., 2013). However, tumours exhibit histological and molecular heterogeneity. A tumour is not a homogeneous population of malignant cells, it is a heterogeneous population of cells where the sub-clonal populations best adapted to their environment gain a selective advantage and proliferate (Stratton et al., 2009; Vogelstein et al., 2013). The evolutionary model of the tumour is well established, but it is only recently that researchers have understood aspects of intra-tumour heterogeneity.

Cytogenetic studies of solid tumours normally show multiple tumour cells with differing karyotypes (Höglund et al., 2002). Tannock (2014) found that when multiple regions of single solid tumours were sequenced, between 63 percent and 69 percent of somatic mutations were not detectable in all tumour regions. In addition to evidence for the divergent evolution of tumours there is also evidence for the convergent evolution of subclones, and metastases that independently acquire mutations in the same driver genes (Tannock, 2014). There were also differences in prognostic gene expression signatures across regions of the same solid tumour.

1.4 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) project began in 2006 (Garraway & Lander, 2013). Its aim was to obtain the molecular profile of three cancers (Hudson et al., 2010); glioblastoma multiforme (GBM), serous cystadenocarcinoma of the ovary, and lung squamous cell carcinoma. Since then its scope has increased to include 34 distinct cancer types across 6 data modalities including exome sequence data. The Cancer Genome Atlas adopted the same data standards as the International Cancer Genome Consortium (ICGC) to insure comparability of data across cancer sequencing studies. Despite the precautions taken to standardise the clinical features and molecular data, analysis across cancer types was not trivial. In 2013 the first twelve TCGA cancers were re-analysed using standardised analysis pipelines as part of The Pan-Cancer analysis project (Weinstein et al., 2013). In Chapter 2 I performed exploratory analysis on TCGA and Pan-Cancer colorectal cancer exome sequence datasets to understand the technical effects within the data.

1.5 The Pan-Cancer analysis project

The goal of the Pan-Cancer analysis project was to adopt a coordinated approach to TCGA datasets across twelve tumour types and to apply standardised approaches to analyse the molecular data (Weinstein et al., 2013). Even given the steps taken to standardise the data across the cancer types, some challenges with unknown consequences remained. Challenges to standardising the data included the use of improved array technologies for more recently analysed cancers and multiple sequencing technologies used within each cancer type, resulting in the potential for large batch effects (Weinstein et al., 2013). Each cancer also had a different staging and grading strategy, which made cross-cancer analyses of the most important prognostic indicators non-trivial. The Pan-Cancer analysis project was further developed in 2014, when whole-genome sequencing of Pan-Cancer samples was conducted to supplement the earlier TCGA exome sequencing (Hoadley et al., 2014).

Initial analysis by Kandoth et al. (2013a) found that *TP53*, mutated in 42 percent of

samples, was the most frequently mutated gene across the Pan-Cancer dataset. *PIK3CA* was the second most frequently mutated gene, and was never co-mutated with *PIK3R1*. However, across cancer types there was significant heterogeneity in both *TP53* and *PIK3CA* mutation frequencies. By using the whole-genome sequenced samples and unsupervised clustering techniques, the cancers grouped according to their tissues of origin (Hoadley et al., 2014).

1.6 Cancer staging and tumour grading

The Pan-Cancer data contains cancer grade and stage clinical information. In Chapter 3 I investigated whether there were mutated genes that were associated with cancer grade and stage across cancers.

Cancer prognosis is primarily determined by two factors; the cancer stage, and the primary tumour grade (Engers, 2007). Cancer staging is a relatively well standardised procedure, agreed by the WHO (Edge & Compton, 2010). Staging consists of the measurement of three features, the TNM status. T corresponds to the size of the primary tumour and extent within the tissue, or organ of origin. N denotes the degree of spread to lymph nodes, either local to the tissue of origin, or more distant. The M parameter indicates the presence of distant pathological spread of the cancer where metastasis is either absent or present. By using these three parameters, pathologists and clinical teams establish the clinical stage of a cancer as one of four high level categories of stage I, II, III, or IV, with prognosis becoming successively poorer going from stage I to stage IV. Tumours that have not spread beyond the primary tissue of origin are categorised as stage I, or II. Tumours that have spread to lymph nodes, but not to any other tissues, are categorised as stage III, and tumours that have spread to other tissues are assigned the stage IV classification.

In addition to cancer staging, tumour grading is the next most important prognostic indicator of patient survival (Hammond et al., 2000), and is used by clinicians to inform patient care (Engers, 2007). In contrast to cancer staging which is established using standard guidelines across cancer types, there are no such guidelines to establish tumour grade across tumours of different tissue types. Each cancer has associated with it one, or more, grading

systems which may be used; Prostate cancers use Gleason grading (Epstein, 2010), renal cancers commonly use the Fuhrman system (Fuhrman, Susan A and Lasky, Larry C and Limas, 1982), and ovarian cancers and endometrial cancers mainly use the FIGO grading system (Shepherd, 1989).

Tumour grading is usually conducted by a pathologist (as part of a clinical team) who inspects a slide preparation derived from a tissue biopsy, or tumour resection, using a microscope. The pathologist assesses the degree to which the structure of the tumour is well or poorly differentiated in comparison to the healthy tissue (Epstein, 2010).

Cancer grading systems usually use a three tier, or four tier system. Across all grading systems the scale of tumour differentiation ranges from well differentiated as the lowest number, to poorly differentiated as the highest number. Tumour grading is a subjective process and inter-pathologist agreement for four tier systems is fair at best (Lang et al., 2005; Han et al., 2013; Scholten et al., 2004). By using a two tier grading system (that combines grade I and II in to low grade, and grade III and IV in to high grade) inter-pathologist agreement is better (Kapucuoglu et al., 2008), and the two tier grading systems retain the prognostic power of the four tier systems (Scholten et al., 2004; Kapucuoglu et al., 2008).

At present, the limiting factor to creating a standardised grading system across cancers is the lack of known biological correlates of tumour grade across cancers. In Chapter 3 I investigated whether there are mutational correlates of tumour grade across cancers, where the grading outcome was based on a two tier system.

1.7 High-order cancer types in the Pan-Cancer dataset

The Cancer Genome Atlas and Pan-Cancer analysis projects have used genomic information to discover signatures of various cancer types, with little attention paid to the differences between high-order classes of cancer. The Pan-Cancer dataset is composed of five high-order differentiation subtype classes; adenocarcinomas, squamous cell carcinomas, urothelial carcinomas, blood cancers, and central nervous system cancers.

All carcinomas originate from the epithelial tissues such as the skin, or those which line

organs such as the colon. There are various subclasses of carcinomas defined by the specialised functions of the epithelial cells. Adenocarcinomas originate from glandular epithelial cells. Squamous cell carcinomas originate from the squamous epithelial cells on the surface of the body, or tissue lumen. Urothelial carcinomas begin from a type of epithelium specific to the urinary tract and bladder. The blood cancers, such as acute myeloid leukaemia (AML) originate in the bone marrow tissue. Cancers of the central nervous system originate in the spinal cord, or brain.

The cell types of each of these broad cancer categories perform specialised functions and have diverse morphology. It is conceivable that the mutational processes that contribute to cancer formation and progression differ between each high-order subtype of cancer. Weinstein et al. (2013) found that Pan-Cancer types generally clustered according to their tissue of origin, and that the lung squamous cell, and head and neck squamous cell carcinomas were grouped. By grouping the twelve Pan-Cancer cancer types into five high-order categories the patterns of mutated genes that discriminate between the cancer types may be identified and new disease mechanisms revealed. I investigate how the five categories differ in their patterns of somatic mutation in Chapter 4.1 on page 101.

1.7.1 Cancers of unknown primary origin

Around three percent of all diagnosed cancers are cancers of unknown primary (CUP) origin. These are stage IV metastatic cancers that usually occur at multiple sites across the body. Prognosis for these cancers is typically poor. A correct diagnosis of the primary tumour may have implications for patient care. However, identification of a primary tumour is difficult, and is most often possible post mortem. Gene expression tools have been developed to predict the primary origin of CUP (Tothill et al., 2005; Ramaswamy et al., 2001; Su et al., 2001) , but no tool has been developed to predict CUP origin based on whole exome sequence data. Metastases are thought to inherit the mutations that occurred in the primary tumour. Due to the successful discrimination of high order cancer types in Chapter 4.1 a Random Forest approach was used in Chapter 4.2 to generate a model to assign the Pan-Cancer

cancers to their tissue of origin based on exome sequence data. Genomics England will conduct whole genome sequencing on CUPs, thereby generating the data to validate this Random Forest approach.

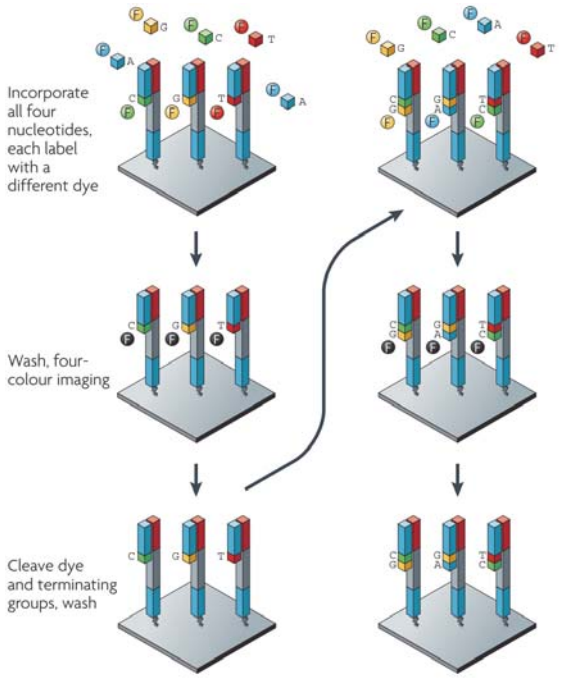
1.8 An outline of exome sequencing

Within the last decade, a new generation of sequencing technologies have been developed which allow the fast parallel sequencing of millions of reads of DNA. One of the most popular NGS technologies is the Illumina sequencing by synthesis technology.

The human genome is composed of around 3 billion nucleotides, whereas the exome is approximately one percent of that size (30 million nucleotides). The exome contains all of the DNA sequence that is transcribed in to mRNA and eventually translated in to protein. As proteins constitute the majority of the cellular machinery, mutations in proteins are expected to be most likely to result in a disease. Whole exome sequencing was undertaken by TCGA to study the DNA alterations that cause cancers.

In the case of exome sequencing the first step is exon capture, using one of the exon capture kits. The DNA library is then prepared on a *flow cell* using solid-phase bridge amplification, a modified version of polymerase chain reaction (PCR), to create millions of clusters of clonal reads of up to 150 base-pairs in length (Metzker, 2010; Berglund et al., 2011). Sequencing-by synthesis (SBS) is then performed to characterise the sequence of upwards of hundreds of millions of short sequences simultaneously (Figure 1.8.1). Each cluster of template sequences should be identical and have incorporated the same dye modified nucleotide during each round of SBS. The nucleotide incorporated at each step is inferred from the colour of fluorescent light emitted by the cluster upon excitation with a laser. In this way, each round of SBS represents the addition of a single nucleotide, whose identity is known by the colour of emitted light. For each cluster of templates, the nucleotide sequence is inferred from the order of colours recorded from each successive round of SBS (Metzker, 2010).

Figure 1.8.1: Sequencing by synthesis



DNA polymerase is bound to the primer templates and a single dye labelled nucleotide is incorporated in to far end of the strand. The DNA polymerase reaction is halted because of an inhibitory group attached to the incorporated nucleotide. All nucleotides are then washed off the flow cell and imaging of the entire flow cell is performed. The dye group and polymerase inhibiting group are then removed from each template in a cleavage step, and the next addition of dye modified nucleotides begins. The figure is reproduced from Metzker (2010) with copyright permission (License Number 3707211251218).

1.8.1 Types of mutations uncovered by exome sequencing

Exome sequencing is used to identify mutations, such as single nucleotide variants (SNVs) and small insertions and deletions (indels), that occur in the protein coding sections of genes and 5' and 3' un-translated regions. Synonymous SNVs do not cause a change to the amino acid sequence of the protein encoded by the gene. However, non-synonymous SNVs cause a change in the amino-acid sequence that may affect protein function. Non-synonymous SNVs also include premature start and stop codon mutations, which change the positions in the mRNA transcript at which mRNA is translated in to protein. Start and stop codon mutations

can have catastrophic effects on protein function by prematurely initiating translation of the mRNA molecule, or truncating the translated peptide chain respectively. Both mutations have potential to severely affect the tertiary, three-dimensional, structure of the protein and its function.

Indels are typically rarer than SNVs. Indels of single nucleotides result in frame-shift mutations where the triplet amino acid code is disrupted, that is, all the amino acids following the indel would be modified. The resulting protein may not function normally. Indel mutations tend to occur in multiples of three, such that the post-indel amino acid sequence remains intact.

Exome sequencing can also detect copy number variants (CNVs): sections of DNA that are repeated a certain number of times. Chromosomal rearrangements may also be detected using exome sequence data where two sections of the same sequence read map to different chromosomes.

1.8.2 Exome-sequencing in cancer studies

The reduction in the price of NGS technology has brought with it the ability to sequence the genomes of thousands of individuals. Cancerous tumours harbour both the mutations that are unique to the cancer (somatic mutations), and also the mutations that were inherited from parents (germ-line mutations). Exome sequencing and variant annotation must be conducted on the normal tissue and tumour tissue to identify the somatic mutations. The germ-line mutations that occur in both the normal tissue and the tumour are discarded to leave the somatic mutations.

Exome sequencing can give unparalleled characterisation of the mutation profile of a human tissue sample, but only in the coding regions of the genome. The exome comprises around one percent of the human genome, leaving 99 percent completely uncharacterised. Vogelstein et al. (2013) suggested that based on exome studies of monogenic diseases, we can expect to uncover around 80 percent of the variants contributing to cancer by focusing on coding regions. These regions are more likely to affect function of the resultant protein

and contribute to cancer than a non-coding mutation (Vogelstein et al., 2013).

1.9 Clinical practice and next-generation sequencing

As the cost of next generation sequencing is decreasing, clinics and laboratories are considering how to include NGS technology as part of their routine procedures (Meldrum et al., 2011; Garraway & Lander, 2013). For cancers this may mean using the mutational profile of a tumour to make predictions about current clinical measures and outcomes. The task of cancer bioinformaticians is to provide relevant information about mutations associated with clinical features such as cancer sub-type classifications, or cancer grading and staging.

Recent studies have used exome-sequence panels to identify mutations that can guide patient care, most notably in terms of targeted therapy (Frampton et al., 2013). Already, copy number variants (CNVs), SNVs, and small indels can be genotyped using NGS technology and used identify the genes which can guide targeted therapies for individual patients (Frampton et al., 2013).

Cancer diagnostic tests based upon sequence mutations are currently in development. In oesophageal cancer, a potential diagnostic test that uses a combination of Cytosponge sample collection and whole-genome sequencing has discriminated between pre-malignant adenomas and adenocarcinomas (Weaver et al., 2014). Tests that analyse circulating-tumour DNA sequence show potential for real-time monitoring of the effects of cancer treatment and disease progression using blood plasma samples (Dawson et al., 2013; Newman et al., 2014; Diaz & Bardelli, 2014). For colorectal cancer, a test that measures epigenetic modification in cells collected from fecal samples has had success (Ned et al., 2011).

As sequencing technology becomes cheaper, it will become viable for clinics to move most cancer testing over to NGS methods because a single sequence run may be used to conduct many tests (Rehm, 2013; Garraway & Lander, 2013). The Genomics England project is focused on integrating NGS analysis into standard NHS practice by performing whole genome sequencing on 100 000 genomes. The Genomics England Clinical Interpretation Partnership (GeCIP) composed of NHS and academic researchers will collaborate with experts from

private companies to "...accelerate the development of new diagnostics and treatments for NHS patients..." (Genomics England, 2015). The GeCIP will also feedback findings for individual patients to clinicians to inform treatment decisions (Caulfield et al., 2015).

1.10 An overview of machine learning

This thesis makes use of Random Forest and other machine learning methods. I will now provide an overview of machine learning methods and Random Forest in particular.

The amount of genomic data available on the world wide web is increasing. In the case of TCGA, upwards of 5000 whole exome sequence datasets are available to researchers. There is a corresponding increase in information in the available genomic datasets, which provides opportunities for bioinformatics research. Apart from classical statistical techniques, algorithms can be used to extract the information within the data by pattern recognition. The development of algorithms for learning these patterns is a branch of computer science called machine learning.

In bioinformatics research machine learning methods can be used to develop patterns and rules that describe a well characterised set of samples. The rules can be used to classify new, unseen, samples. Complex interactions between multiple features may not be appropriately described using classical statistical methods and may not be easily included in statistical prediction models. However, machine learning methods offer a solution to include complex interactions between features in a prediction model.

1.10.1 Supervised and unsupervised methods

Supervised learning techniques use a set of samples for which the outcome class is known. Rules are discovered that describe each class, and the rules are used to classify samples whose class is unknown. Unsupervised learning techniques take a set of samples without class labels and identify clusters of samples based upon the features. The clusters are then described based on patterns in the features.

1.10.2 Evaluation of machine learning methods

Machine learning methods for classification tasks are heavily reliant on the dataset being an accurate representation of the population. Random noise in the dataset used for model

building may lead to a classification model that exploits attributes that discriminate only between the samples of that specific dataset, resulting in an *over-fitted* model that does not perform well when applied to new data. There are numerous methods employed by machine learning researchers to mitigate *over-fitting*, and measure any over-fitting which may be present in a classification model.

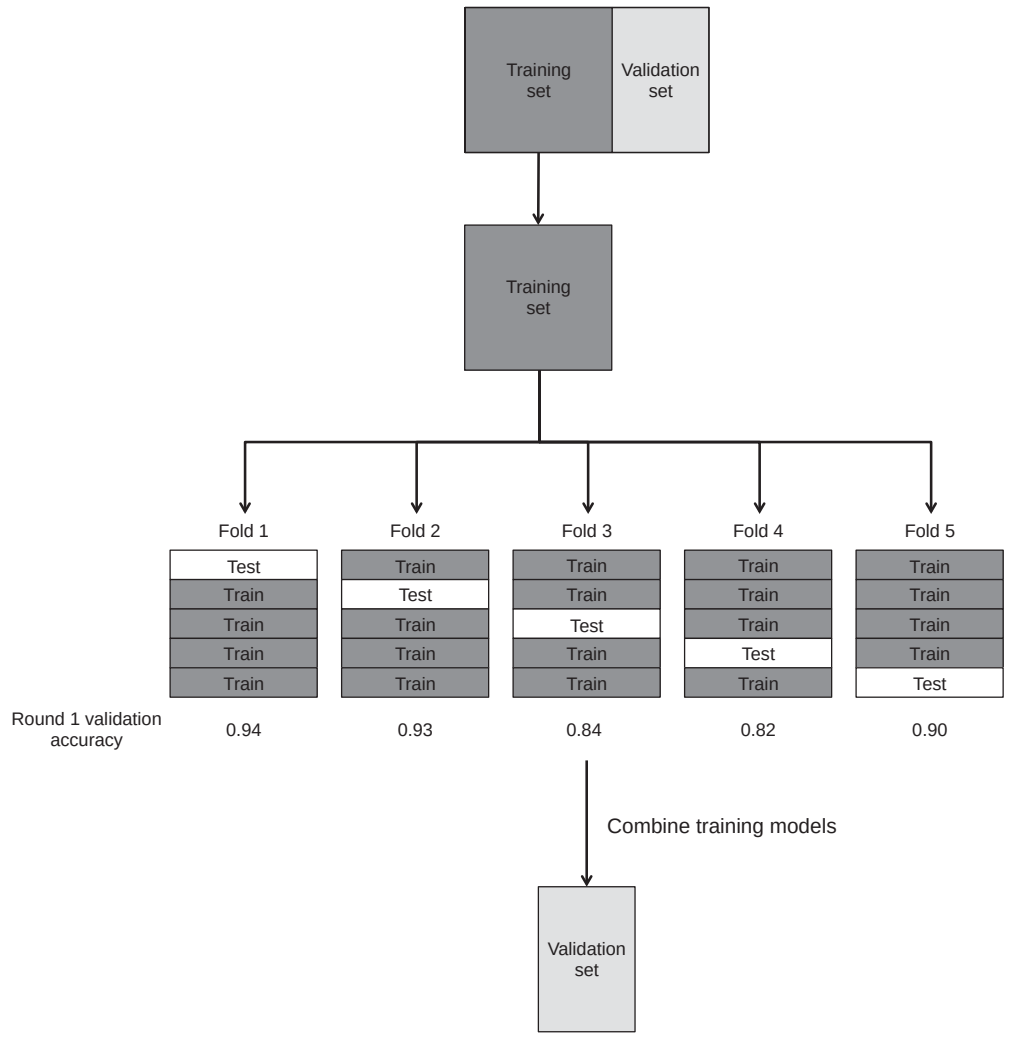
1.10.2.1 Independent test data

Machine learning classifiers are usually built using a subset of the whole dataset known as a *training set*. Using a training set to measure the overall performance of a classifier leads to over-optimistic estimates of performance because of over-fitting. The performance of the model can then be measured in the part of the dataset that was not used for model building, known as the *test set*. If a dataset is large, partitioning it in to a training set and test set is desirable.

1.10.2.2 K-fold cross-validation

When using datasets that have been partitioned in to a training set and a test set, uninformative patterns among features in the training set can lead to an overfitted classification model. *K-fold cross-validation* (Stone, 1978) provides a way to further estimate the extent of the over-fitting. The entire dataset is first randomly split in to k number of *fold*s. Each fold is treated as a test set in turn, the remaining k-1 folds are treated as the training set. Classifier performance is measured as the mean classification accuracy across the k folds. In order to maintain a similar class composition in each fold as the overall dataset, stratified cross-validation can be used. In stratified cross-validation each fold is composed of a similar proportion of each outcome class as the training set. A typical approach is to repeat the cross validation procedure five times, each time having randomly selected samples for each fold (Figure 1.10.1). The cross-validation steps can be repeated to mitigate bias in the way the cross validation folds have been defined.

Figure 1.10.1: k-fold cross validation procedure



A depiction of the cross validation process where a data set is first split in to a training set for model building and a validation dataset. The training set is split in to five folds. Each fold is used as the test set in turn while the remaining sets are used to train the model. Each round of model building has associated with it a classification accuracy. The mean classification accuracy across the five rounds of model building provides an indication of the extent to which overfitting may be a problem with the dataset. The models from the folds are usually combined in some way, trained on the entire training set, and then tested in the validation set. The validation set accuracy can be expected to be achieved in unseen data.

1.10.2.3 Validation datasets

If cross-validation techniques are applied to the entire dataset during model building, no test set will be available to assess the classification accuracy of the model. Each fold will have been used for model construction and model testing. An independent *validation* dataset is required to obtain an unbiased estimate of how the classifier will perform on future datasets (Figure 1.10.1).

1.10.2.4 Over-sampling and under-sampling techniques

When outcome classes are imbalanced, a binary classifier can tend toward choosing the majority classification for all samples (Barandela & Sã, 2003). In order to balance outcome classes two main techniques can be used: *up-sampling*, and *down-sampling*. Down-sampling can be used to reduce the size of the majority class down to the size of the minority class, and can be thought of as jack-knife (Efron & Gong, 1983) sampling the majority class. However, this comes at the cost of removing information contained in the unused majority class samples from the model building process. Up-sampling, or sampling with replacement, can be used to increase the size of the minority class to equal that of the majority class. Bootstrapping (Efron & Tibshirani, 1985) is commonly used to randomly duplicate samples of the minority class. The synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) is more sophisticated. SMOTE nominates a *seed sample*, and creates new synthetic samples by combining features from the seed sample's k-nearest neighbours.The SMOTE algorithm is used in Chapter 4.2 to increase the size of a small training set.

1.10.2.5 Classification statistics and visualisation

The results of machine learning tasks are commonly presented as classification statistics. For binary classification tasks the terms *sensitivity* or recall, and *specificity* or precision, are typically used to measure the performance of classification models (Table 1.10.1).

Table 1.10.1: Binary classification statistics

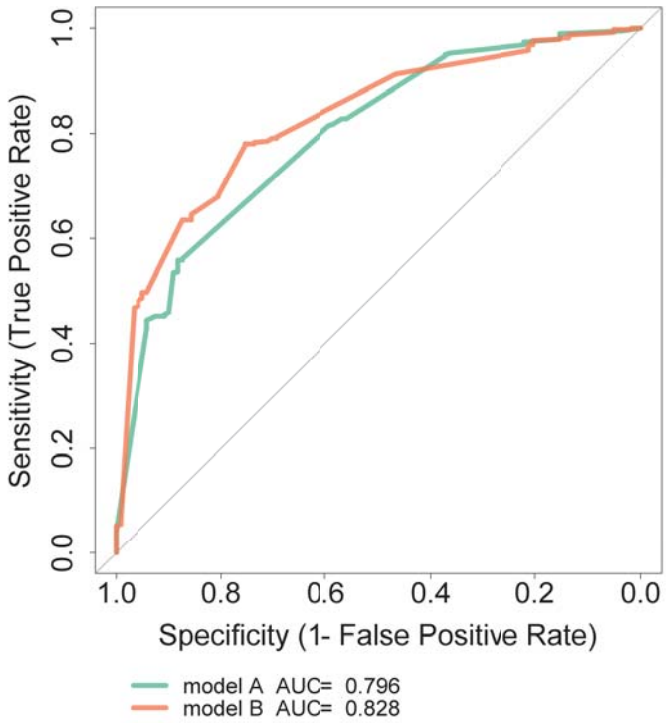
		Condition	
		Condition positive (CP)	**Condition negative (CN)**
Test Outcome	**Test outcome positive (TOP)**	True Positive (TP)	False Positive (FP)
	Test outcome negative (TON)	False Negative (FN)	True Negative (TN)

$Specificity = TN/CN$ $Negative\ Predictive\ Value = TN/TON$
$Sensitivity = TP/CP$ $Positive\ Predictive\ Value = TP/TOP$
$Accuracy = TP + TN/TP + FP + FN + TN$ $Prevalence = CP/CP + CN$

The calculation of classification statistics are shown in relation to the contingency table for the possible classifications during a binary classification task.

Receiver Operator Characteristic (ROC) curves are commonly used to visualise how the sensitivity (true positive rate) and specificity (1- false positive rate) change across classification model thresholds (Figure 38). A plot lying above the diagonal indicates that the classifier performed better than chance. ROC curves provide a way to compare classification models based on the integral of the ROC curves, or Area Under the Curve (AUC) statistic. The AUC is a summary of how a classifier performs at all possible classification threshold probabilities. Classification models with higher AUC are generally preferable to those with a comparatively lower AUC. ROC curves and the AUC statistic are used in two chapters in this thesis (Chapters 3, and 4) to visualise classification results.

Figure 1.10.2: ROC curve example



Two ROC curves generated for models A and B. Model B has a higher AUC than model A and may be considered a superior classification model. The diagonal line indicates the plot that would be generated by a non-informative classifier.

1.10.3 Decision tree classifiers

Decision Trees are supervised classifiers composed of a set of decisions represented as a *tree*. In a tree every decision is represented by a node, and each node that has daughter nodes indicates a decision to be made on the values of a feature. The C4.5 algorithm (Quinlan, 1993) is a popular single tree classification method. It provides a good example for the construction of single trees based upon a dataset consisting of N samples, M features, and one binary outcome O. At each decision the information gain achieved by partitioning samples based on each feature is measured. Information gain can be described as the increase in enrichment of outcome classes among subclasses in comparison to the parent class. The attribute m_j with the highest information gain is used to partition the samples in to two subsets. This process continues recursively until all subclasses consist of a single outcome

class. These terminal classes can be described as *leaves*. Some decision tree algorithms *prune* back leaves of each tree in order to mitigate over-fitting.

1.10.4 Random Forest

In Chapter 4.1 I have used Random Forest to identify the features that discriminate between five classes of cancer based upon the somatic exome mutation profile of the cancers. A Random Forest (Breiman, 2001) is a supervised ensemble classifier based upon multiple decision *trees*. Each decision tree in the group of trees, or forest, generates a class prediction for each sample. For each sample the class prediction of the forest is the most common class predicted across all trees in the forest.

Random Forest uses two techniques to avoid overfitting: *bagging*, and *random feature selection*. In a Random Forest analysis a training set and test set can be used where model building is conducted on the training set. The training set is divided in to a *learning set* and an *out-of-bag* set. For each tree a bootstrap sample of the learning set is used. A tree is learned by splitting the set of samples based on the feature that best partitions the samples. This process continues until each subset of samples after a decision are of the same class. The out-of-bag set is used to measure the classification accuracy of the forest, and the importance of each feature. The out-of-bag samples are run through each tree in the forest.

The importance of each feature is commonly measured as either the gini impurity (Kuhn & Mori, 1995), or mean clarification accuracy (Breiman, 2001). These measures provide a summary of the ability of each feature to correctly classify the out-of-bag set across the forest.

In order to prevent over-fitting due to out-of-bag and learning set definition, cross-validation approaches can be used. In addition to the bagging approach, Random Forest also randomises the set of features from which a decision can be made at each node of each tree.

1.11 How protein interaction networks can be used to identify new disease gene candidates.

Chapter 5 of this thesis uses protein interaction networks to prioritise candidate disease genes by supplementing gene-level scores with protein interaction data. I will now provide an overview of protein interaction networks, some background on graph theory, the network properties of disease genes, and a summary of the methods used to prioritise candidate disease genes using network data.

As mentioned in section 1.3.2.1 on page 22 genetic heterogeneity may prevent some of the genes that contribute to cancer from being discovered using gene-level tests. Univariate tests often assume independence between the features being tested. For genes, the assumption of independence is not appropriate. Genes code for proteins, and the function of a protein may be a single step in a larger pathway composed of a group of proteins. Across a sample of individuals, disruption to the function of a different protein from the same pathway in each sample may cause the same disease. For each univariate test conducted, only a subset of cases will carry disease variants in a gene, and other cases will carry mutations in other genes in the same pathway (Leiserson et al., 2013b). Protein-Protein interaction (PPI) networks contain information about which proteins physically interact and may be part of a pathway. By combining the analysis of univariate data with the information within protein interaction networks researchers may identify the pathways that are enriched for genes associated with diseases and possibly untangle some of the genetic heterogeneity of a disease.

1.11.1 Protein-protein interaction networks

Two main high-throughput technologies have been important for the creation of PPI networks; Yeast-2-Hybrid (Y2H) (Young, 1998), and tandem affinity purification and mass spectrometry (TAP-MS) (Gavin et al., 2002). The Y2H system is used to test for binary protein interactions across the proteome in a pairwise manner. The TAP-MS technique is able to identify protein complexes consisting of more than two proteins, by first 'pulling down' the protein complex

and then by characterising the constituent proteins using mass spectrometry techniques.

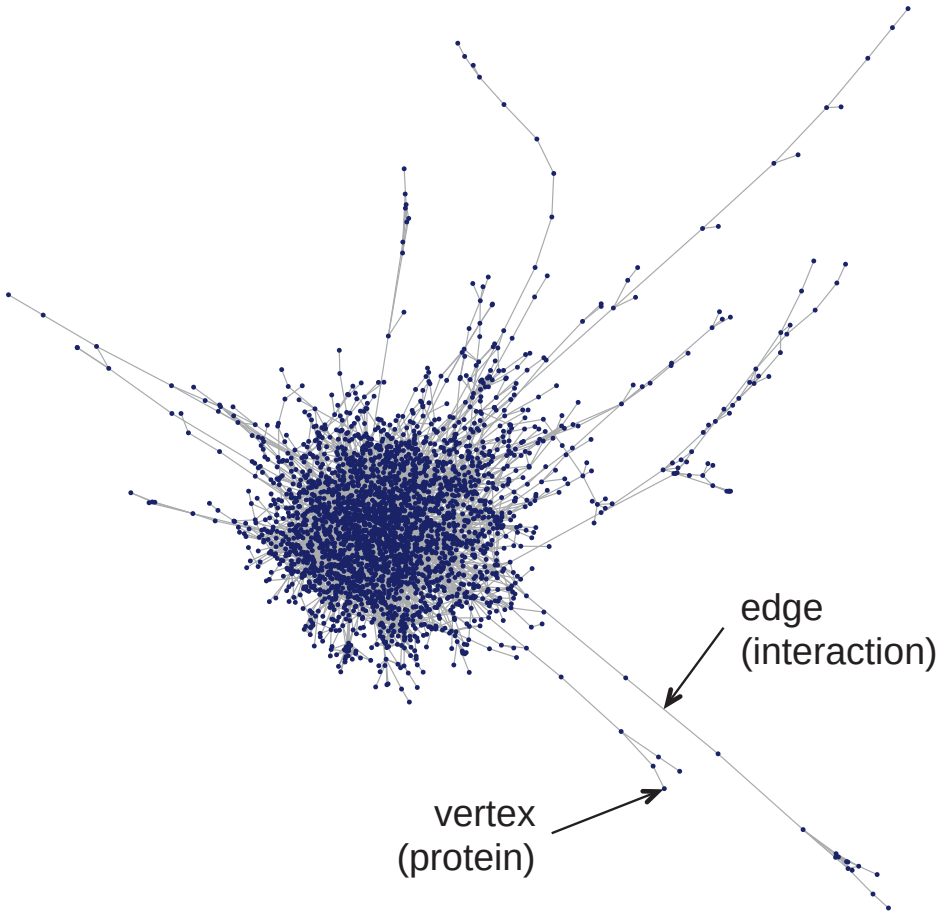
Various protein interaction databases have been created by searching the literature for reports of protein interactions in small experiments. There are now many available databases, including; the Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009), the Molecular INTeractions database (MINT) (Licata et al., 2012), the Biological General Repository for Interaction Datasets (BIOGRID) (Stark et al., 2006), the Biomolecular Interaction Network Database (BIND) (Bader et al., 2003), and the IntAct molecular interaction database (Hermjakob et al., 2004). Other databases have been created by combining much of the information available in these public resources, including; the irefweb (Turner et al., 2010) database, STRING (Franceschini et al., 2013), the protein interaction network analysis platform (PINA) (Cowley et al., 2012), the Human Protein Protein Interaction network (HuPPI) (Lehne & Schlitt, 2009), and the Consensus Pathway Database (CPDB) (Kamburov et al., 2013). The HumanNet (Lee et al., 2011), and PrePPI (Zhang et al., 2012) databases also provide predicted protein interactions.

1.11.2 An overview of graph theory

Networks can be represented as graphs to allow analysis. A PPI network can be represented as a graph $G = (V, E)$, where proteins are represented by vertices V, and interactions between the proteins are represented by edges E (Figure 1.11.1). A *directed graph* describes relationships of cause and effect where information flows from one vertex to another: each edge being assigned a direction to indicate information flow. Gene regulation networks are examples of directed graphs. A graph that imposes no restriction on the direction of information flow across edges is called an *undirected graph*. PPI networks can be described as undirected graphs. Throughout this thesis, only undirected graph representations of PPI networks are used.

Graphs have many numeric properties which can be exploited for analysis and all can be broadly categorised as either a local property, or a global property.

Figure 1.11.1: Protein interaction network graph



The largest component of the HuPPI2 (Lehne & Schlitt, 2009) human protein protein interaction network. The small blue circles are vertices that represent proteins, and the grey lines that connect the vertices are the edges, or physical interactions between proteins.

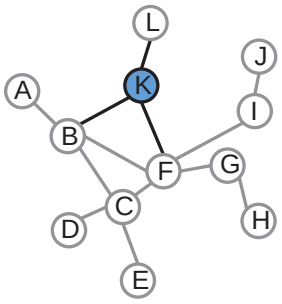
1.11.2.1 Vertex measures

Vertex measures are local graph properties. Each vertex can be described by a number of vertex properties, the most simple of which is vertex degree (Figure 1.11.2a). The vertex degree is defined as the number of edges that connect to a given vertex. Clustering coefficient (Figure 1.11.2b) indicates the density of edges among a vertex's adjacent vertices as the proportion of edges among the adjacent vertices in comparison to the total possible number of edges. Assortativity (Figure 1.11.2c) describes the mean degree of the vertices adjacent to a vertex. Other metrics describe each vertex in relation to the entire network, and not just

its local neighbourhood. The concept of a path, the series of edges between any two given vertices, can be summarised as a path length. The shortest path length (Figure 1.11.2d) is important for the calculation of other network properties. The betweenness centrality (Figure 1.11.2e) of a vertex describes the importance of each vertex for efficient information flow through the graph as the number of shortest paths that pass through the vertex.

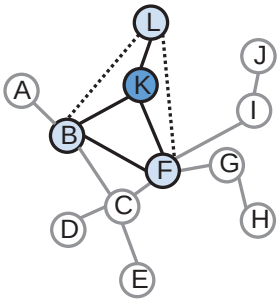
Figure 1.11.2: Vertex measures

(a) Degree



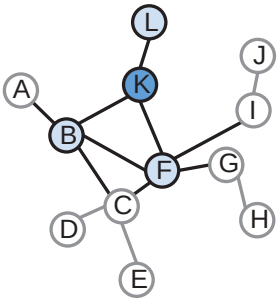
$M_K = 3$

(b) Clustering Coefficient



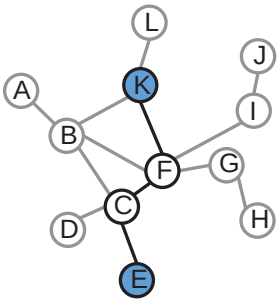
$C_K = 2n_k / M_K(M_K - 1)$

(c) Assortativity



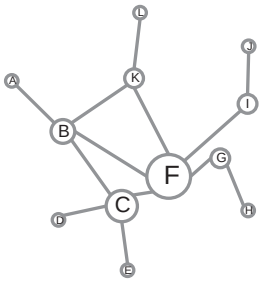
$NC_K = \left(M_B + M_F + M_L\right) / M_K$

(d) Shortest Path



$SP_{EK} = (E, C, F, K)$

(e) Betweenness



$B_K = 0.18$

$B_F = 0.64$

The importance of a vertex and its local structure in a graph can be characterised by several measures. (a) The Degree (M_K) is the number of edges that vertex K shares with other vertices. (b) The Clustering Coefficient describes the interconnectivity between the neighbouring vertices of K and is defined as the number of edges between neighbouring vertices of K (n_K) divided by the number of possible edges between neighbours of K. (c) The Assortativity NC_K is defined as the average degree of the neighbours of K. (d) The Shortest Path SP_{EK} between vertices E and K is defined as the smallest number of edges between the pair of vertices. (e) The Betweeness B_K of vertex K is the number of shortest paths in the graph which pass through vertex K. Figure adapted from Yamada & Bork (2009).

1.11.2.2 Global graph properties

A graph can be described using a number of global properties. The number of vertices and edges can give an indication of the size and density of a graph. The diameter of a graph, or the mean shortest path length, indicates the ease at which information can flow across it. Most biological and *real world* graphs, such as telecommunication networks and social networks (Albert & Barabasi, 2002) have a diameter of five, showing that most vertices can reach most other vertices by traversing no more than five edges.

Most biological networks exhibit a *scale-free* topology (Albert & Barabasi, 2002; Ravasz et al., 2002). The degree distribution of a biological graph will typically follow a power-law distribution, with many vertices having a small degree and very few vertices having a large degree.

1.11.3 Network properties of disease genes

Proteins that have a high degree are highly connected and are known as *hubs*. Hubs usually correspond to genes that perform essential functions in the cell (Jeong et al., 2001). Mutations in hub genes tend to be lethal and evolution is more constrained in these genes (Saeed & Deane, 2006). There is however limited consensus on the subject. Wachi et al. (2005) found that disease genes tend to be hubs. However, Goh et al. (2007) found that when disease genes that are also essential are removed from a PPI network, the remaining disease genes are not more highly connected than other genes. Genes for monogenic disorders are central to protein interaction networks, whereas genes contributing to complex diseases are closer to the periphery, although they are closer to one another than expected (Barrenas et al., 2009). Non-essential disease genes tend not to be hubs, but are more central to the PPI network than expected. Proteins that are associated with the same disease are ten times more likely to interact physically (Goh et al., 2007). Proteins that have a similar biological function as defined by Gene Ontology (GO) terms tend to be in close proximity in a PPI. Similarly, proteins that are involved in the same or similar diseases tend to cluster in protein interaction networks (Goh et al., 2007; Gandhi et al., 2006). These properties of disease

proteins in general indicate that a disease can be associated with a community, or module, of vertices in a PPI network. A community can be a group of highly interconnected group of vertices.

There is heterogeneity across cancers in terms of the commonly mutated genes. Even within cancers, there is no single gene which is mutated in all samples (Vogelstein et al., 2013). The originating mutation of cancers of the same type can be heterogeneous. The apparent heterogeneous driver mutations of cancers of the same type may be due to mutations in functionally related genes which are close, or adjacent in a protein interaction network (Oliver, 2000). The proteins that have not been associated with the disease, but that interact closely with disease causing genes would be potential candidate disease genes. This is an example of *guilt by association*. The guilt by association hypothesis is central to many of the methods that prioritise and predict new disease genes. Chapter 5.1 on page 153 of this thesis describes the development of a network-based method for disease gene prioritisation based on the guilt-by-association hypothesis.

1.11.4 Predicting disease genes using networks

Many approaches prioritise candidate genes based on the proximity of candidate genes to disease genes. Candidate genes that are closer to disease genes are ranked more highly (Leiserson et al., 2013b). The simplest methods employ neighbour counting (Oti et al., 2006) to rank candidate genes according to the number of disease genes to which they are adjacent in the PPI network. Two proteins can be involved in the same disease pathway without interacting physically. There are methods that quantify the *closeness* of candidate genes and known disease genes in terms of the shortest path metric. Krauthammer et al. (2004) used a *molecular triangulation* method, which ranked candidate genes with small shortest path distances to known diseases genes most highly as disease gene candidates. In Alzheimer's disease, their highly ranked candidates agreed with a manually curated gene set. The CFinder algorithm (Palla et al., 2005, 2007) identifies structures in the PPI network known as cliques. In a PPI network a clique corresponds to set of proteins where each

member of the set shares an edge with all other members of the set. CFinder then identifies the cliques which intersect by k members and treats the resultant communities as important.

In addition to local network properties, global network properties have been used to prioritise candidate disease genes. The *random walk with restart* (Tong et al., 2008) can successfully identify known disease genes in a network (Köhler et al., 2008). Random walk methods typically perform better than local network property methods (Navlakha & Kingsford, 2010). However, methods that combine the results of local and global network approaches perform best (Navlakha & Kingsford, 2010).

The above methods use only the network structure to predict and prioritise candidate disease genes. There are many *de-novo pathway discovery* methods that integrate protein interaction network data and 'omic experiment data to discover perturbed pathways and disease genes.

1.11.4.1 De-novo pathway discovery methods

The jActiveModules (Ideker et al., 2002) cytoscape (Shannon, 2003) plugin was the first algorithm to search for modules in a PPI network that was integrated with experimental data. It used a greedy simulated annealing (Kirkpatrick, 1984) approach to identify the subnetwork with the highest *activity score*.

Chuang et al. (2007) developed a method that uses gene expression data to discover differentially expressed subnetworks in a PPI network. The method takes a gene-list of differentially expressed genes and *grows* differentially expressed subnetworks that best discriminate between cases and controls using a greedy algorithm. From a differentially expressed *seed* protein, an additional neighbouring protein is added to the group. The mean z-score of the proteins included in the subnetwork is computed for cases and controls, and the additional protein is permanently included in the subnetwork if the discrimination between cases and controls (the mutual information) is increased above a user-defined threshold. The HyperModules (Leung et al., 2014) cytoscape software takes a similar greedy approach. It grows modules from mutated seed genes to uncover the subnetwork that best predicts a

binary clinical outcome, or survival. The DEGAS (Ulitsky et al., 2010) algorithm aims to find a perturbed subnetwork by solving a minimum covering set problem. It uses a greedy approach to identify the smallest subnetwork that covers all but n cases k times each. The authors use the proteins which are differentially expressed in each individual as seeds for the analysis, but any gene-based test for which each sample can be given a binary score at each gene could be used. The NetBox (Cerami et al., 2010) algorithm identifies modules of mutated genes that are either adjacent in a PPI network, or are connected through a single linking gene. KeyPathwayMiner (Alcaraz et al., 2011) identifies the maximally connected subnetworks in which the number of non-labelled (non-significant) genes does not exceed a user-defined k by using an Ant Colony Optimisation (Dorigo et al., 1999) technique. The Dapple algorithm (Rossin et al., 2011) uses a gene list, protein interaction network data, and recombination hotspot data to test whether genes in the list are more closely connected in the PPI network than would be expected. It has been used to provide candidate genes for genome wide association studies (GWAS), where an associated locus may contain many genes. These methods compute the statistical significance of the resulting subnetworks, by comparison to an empirical distribution of subnetwork scores, achieved in a variety of ways including vertex label swapping, edge swapping, and case-control sample swapping where appropriate.

Network methods are limited by the coverage and quality of the network used. It is unlikely that any of the current human PPI databases are complete. PPI databases contain false positives, which will introduce noise and affect the results of network-based analyses. Most networks are not tissue specific, and may be a result of the union of interactions across many cell types and conditions (Leiserson et al., 2013b).

1.12 Study aims

This study aims to identify the mutated genes which are associated with clinical features in cancer and to develop a method to prioritise candidate disease genes. Machine learning methods are used to identify the genes which when mutated are indicative of cancer grade,

cancer stage, and cancer types. A network-based method is developed to prioritise candidate genes in cancer exome sequence data and a further network-based method is applied to rheumatoid arthritis GWAS data.

1.12.1 Specific aims of the thesis

- Identify mutated genes associated with grade, and stage across cancers

To establish whether there are mutated genes that correlate with a high grade, or grade 3 and 4 classification, across three types of adenocarcinoma. There is a lack of knowledge of the mutations that indicate cancer grade as low, or high, across multiple cancers. I tested whether there are mutated genes associated with cancer grade using logistic regression analysis while adjusting for participant age, gender, and stage. The same method was used to establish the mutated genes associated with cancer stage.

- Identify the mutated genes that discriminate five high-order cancer types (adenocarcinomas, squamous cell carcinomas, urothelial carcinomas, blood, and brain cancers) from one another

To identify the mutated genes that discriminate five high-order cancer types (adenocarcinomas, squamous cell carcinomas, urothelial carcinomas, blood, and brain cancers) from one another. The Cancer Genome Atlas has focused most cross cancer investigations on identifying new cancer subtypes. Relatively little attention has been paid to identifying the mutations that discriminate between known high-order cancer types in the Pan-Cancer dataset. I used ten 2-class Random Forest analyses to uncover the mutated genes that discriminated between these five cancer types in a pairwise manner. I also created a multi-class classifier to discriminate cancers of different tissue types. For cases of advanced cancer of unknown primary origin this may be useful for pinpointing the primary cancer site in order to inform treatment.

- Use prior biological information in the form of a protein interaction network to suggest new complex disease gene candidates

To use prior biological information in the form of a protein interaction network to suggest new complex disease candidate genes. Using TCGA colorectal cancer exome sequence data and the HuPPI2 (Lehne & Schlitt, 2009) protein interaction network I developed a method called k-pseudo cliques analysis. I used the MutSigCV (Lawrence et al., 2013) test to identify significantly mutated protein coding genes in the TCGA colorectal cancer dataset. Then, I used the HuPPI2 (Lehne & Schlitt, 2009) protein interaction network to identify communities of proteins and establish if any of the communities were enriched for genes that were more mutated than expected based on the gene-level test statistic. The utility of the network method was established by comparing KEGG colorectal and cancer pathway enrichment p-values to that of the univariate test.

In addition, Region Growing Analysis was used to analyse rheumatoid arthritis GWAS data in order to prioritise candidate disease genes which did not reach genome-wide significance.

Chapter 2

Exploratory analysis of The Cancer Genome Atlas and Pan-Cancer colorectal cancer datasets

2.1 Introduction

This chapter describes exploratory analyses performed using The Cancer Genome Atlas (TCGA) colorectal cancer exome sequence datasets from June 2012 (Muzny et al., 2012), December 2012, and the Pan-Cancer 2013 analysis (Kandoth et al., 2013a). I demonstrate how a potentially confounding sequencing technology effect in interim TCGA data from the December 2012 data release was identified, and describe its presence in the TCGA June 2012 data and Pan-Cancer 2013 data. I also investigated the potential to discriminate between micro-satellite instability high samples from micro-satellite instability low samples.

It was common for TCGA studies to use different exome sequence analysis pipelines for each cancer type. The breast cancers (Koboldt et al., 2012), ovarian serous carcinoma (Bell et al., 2011), renal clear cell carcinoma (Creighton et al., 2013), and colorectal cancer (Muzny et al., 2012) used various exome sequence capture kits, and sequence analysis pipelines. I applied a variety of exploratory data analysis techniques to TCGA June 2012, TCGA December 2012 and the Pan-Cancer 2013 colorectal cancer exome sequence datasets to arrive at my decision to use the Pan-Cancer 2013 dataset for all analyses of TCGA samples throughout later chapters in this thesis.

2.1.1 Confounding factors

The exome sequencing process relies on a complex set of procedures, including, exon capture, polymerase chain reaction amplification of DNA, the sequencing step itself, and downstream processing of the raw sequence data. When these conditions are varied within an experiment the measured outcomes may vary according to technical factors. Technical effects can often account for a larger proportion of the variance in an experiment's outcome measures than the intended biological effects, thereby calling in to question the biological findings of the experiment.

2.1.1.1 Analysis pipelines

The Cancer Genome Atlas (TCGA) exome sequence dataset for colon and rectal cancer was collected using two different Next Generation Sequencing (NGS) technologies and analysis pipelines (Table 2.1.1): Illumina sequencing by synthesis technology (Bentley et al., 2008) and ABI SOLiD sequencing by ligation technology (Valouev et al., 2008). The two technologies are fundamentally different and each has systematic errors. For both technologies substitution errors are the most common error type (Metzker, 2010), along with an under-representation of AT-rich and GC-rich regions. The Illumina technology was also known to show an increase in SNV substitution frequency where the preceding base was guanine. In addition, different exome capture protocols, and sequence analysis pipelines were used across the two sequencing technologies in TCGA June 2012, and TCGA December 2012 datasets. (Table 2.1.1). Variant annotation was conducted using MuTect (Cibulskis et al., 2013) for the Illumina sequenced samples, and using SAMTools (Li et al., 2009) for the SOLiD samples (Table 2.1.1). Therefore, technical effects due to sequencing technology cannot be disentangled from a sequence analysis pipeline effect. Throughout this chapter I use the term *sequencing technology effect* to refer to this combined sequencing machine and analysis pipeline effect. In contrast, the Pan-Cancer study used a single variant annotation step across both sequencing technologies. They used the TGI Washington University VariantAnnotator, and removed known false positive somatic mutations and germ-line SNPs present in the dbSNP database (Sherry et al., 2001).

Table 2.1.1: Sequencing technology and analysis pipelines in TCGA and Pan-Cancer datasets

Exome sequencing steps and parameters	Illumina	ABI SOLiD	Pan-Cancer Analysis
exome target capture	NimbleGen SeqCap EZ Exome 2.0 Solution Probes	NimbleGen CCDS Solution Probes	As Illumina and SOLiD
	VCRome 2.1 (HGSC design, NimbleGen)		
post capture PCR	14 cycles	12 cycles	As Illumina and SOLiD
Mapping reads	BWA (bwa-0.5.9rcl)	BFAST (version 0.6.4).	As Illumina and SOLiD
HG reference	HG18 BWA reference generated an BROAD and distributed to sequencing centres		HG19
Mutation Detection	MuTect (GATK)	SAMTools Pileup and custom filters	Variant Annotator

Each step going from sequence capture through to variant annotation for the TCGA June 2012, TCGA December 2012, and Pan-Cancer analysis dataset is shown above. In the Pan-Cancer Analysis column, the only modification to the analysis of the Illumina and SOLiD data was the mutation annotation step.

2.1.2 Micro-satellite instability

In colorectal cancer there are two important classes of tumours, the tumours which are hypermutated, and those which are not. Hypermutated tumours carry many SNV mutations, which are often accompanied by a micro-satellite instability high (MSI-H) phenotype. Micro-satellite instability is measured using the Bethesda micro-satellite marker panel consisting of five markers (Umar et al., 2004; Fidalgo et al., 1998; Boland et al., 1998) three of which are dinucleotide markers (D2S123, D5S346, and D17S250), and two are mononucleotide markers (BAT25, and BAT26) (Boland et al., 1998; Dietmaier et al., 1997; Bocker et al., 1997).

Based on the Bethesda panel testing each tumour was classified as one of three types by TCGA: micro-satellite instability stable (MSS), micro-satellite instability low (MSI-L), and micro-satellite instability high (MSI-H) (Umar et al., 2004). A tumour was classified as MSS when there was no difference in the copy number of any of the markers between the normal tissue and tumour tissue. A tumour was classified as MSI-L if a single marker differed in copy number between the normal tissue and tumour tissue. When two or more markers differed in copy number between the normal tissue and tumour tissue, a tumour was classified as MSI-H (Boland et al., 1998; de la Chapelle & Hampel, 2010).

MSI-H occurs in around 15% of colorectal tumours (Boland et al., 1998) when mismatch repair (MMR) pathway genes malfunction. It can be caused by a recessive mechanism when a heterozygote acquires inactivating mutations in the wild-type copy of *MSH2*, *MLH1*, *MSH6*, or *PMS2* (Parsons et al., 1993), or because of two mutations arising independently in both working copies of an MMR gene in a wild-type homozygote (de la Chapelle & Hampel, 2010). The MSS and MSI-H annotations imply fundamentally different routes to cancer.

By using a selection of descriptive, inferential, and exploratory methods I analysed three TCGA data sets (TCGA June 2012, TCGA December 2012, and Pan-Cancer). I investigated potential technical effects using over-representation analyses and principal component analysis (Pearson, 1901), and used non-metric Multi Dimensional Scaling (nmMDS) (Kruskal, 1964) to visualise potential technical effects driving variation within each of the three datasets.

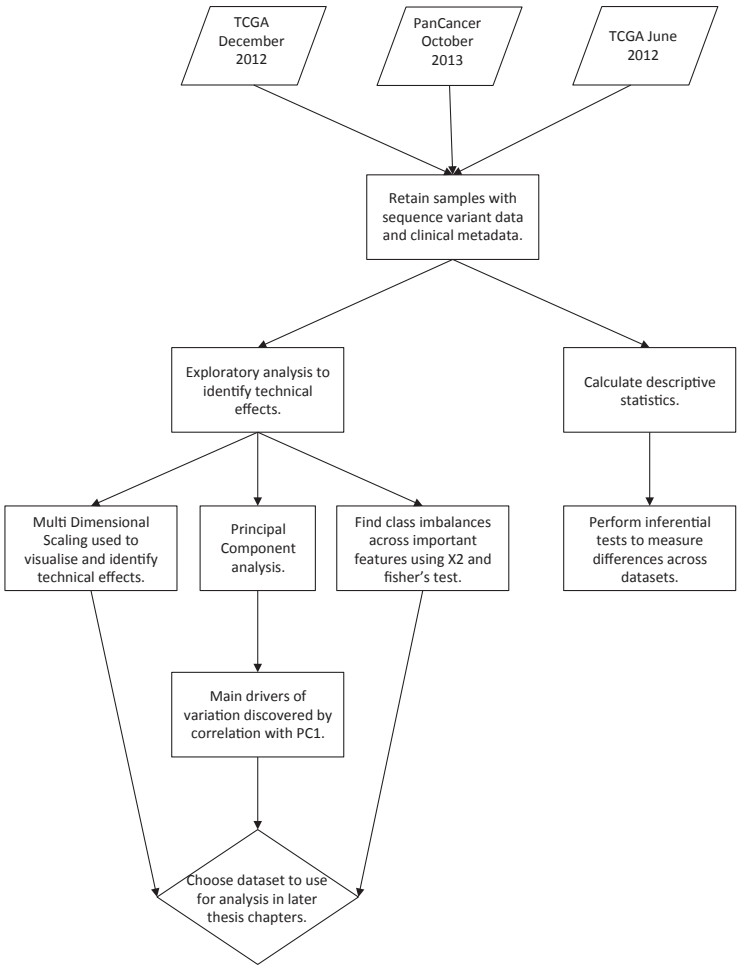
2.2 Methods

I downloaded TCGA December 2012 mutation annotation format (MAF) files and clinical data from TCGA data portal on December 17th 2012. The Pan-Cancer data was downloaded from synapse (accession syn1710680) on October 3rd 2013. I obtained TCGA June 2012 mutation and clinical data from the supplementary data of Muzny et al. (2012) after personal communication from the Pan-Cancer analysis author indicated that the Muzny et al. (2012) colorectal cancer exome sequence data was the most reliable data release.

Figure 2.2.1 shows the pipeline I used to analyse each of the three TCGA datasets (TCGA June 2012, TCGA December 2012, and Pan-Cancer). I retained the samples for which mutation data and clinical data were available. The Pan-Cancer data and TCGA June 2012 data were composed of the same samples, but used a different sequence annotation step (Table 2.1.1), and there were changes to the clinical metadata. TCGA June 2012 metadata was used where missing from the Pan-Cancer dataset. For each of the three datasets I measured the frequency of SNVs, indels, all mutations and functional mutations. Hypermutated samples were defined by an indel:SNV ratio greater than 10%.

For each dataset I conducted three exploratory analyses (Figure 2.2.1): I used non-metric Multi-Dimensional Scaling (nmMDS) (Kruskal, 1964) to investigate and visualise any phenotypic and technical effects in the data; I combined principal component analysis (Pearson, 1901) with various inferential analyses (kruskal-wallis test, spearman's correlation, and wilcoxon ranks sum test) to identify technical effects driving the variation in the data; and I used over-representation analyses to identify imbalances in the study design. The results of the exploratory analyses informed my decision as to which dataset would be used for all further analyses in this thesis.

Figure 2.2.1: Exploratory analysis pipeline



The analysis pipeline for each of the three cancer datasets; TCGA December 2012, Pan-Cancer October 2013, and TCGA June 2012. For each dataset the samples with sequence data and clinical features were retained. The frequency of SNVs and indels were then calculated along with the most frequently mutated genes. The exploratory analyses included: multi-dimensional scaling, to identify and understand clusters of samples; non-parametric methods for class imbalance analyses including chi-square, Fisher's exact test, and Wilcoxon signed-rank test; and principal component analysis to understand the degreee to which variation among samples was driven by latent variables, and to understand the features associated with those latent variables.

2.2.1 Binary mutation matrix

The MAF files contained somatic mutation data: the mutations that were found only in an individual's tumour and not in their normal tissue. 'Silent' synonymous mutations which do not lead to a change in the amino acid sequence in the translated protein were removed from the MAF file. For each dataset I used the union of the functionally mutated proteins across all samples to generate a matrix M of proteins P by samples S. Each element in M, $M[p, s]$, was set to 0 if there was no protein coding mutation for protein p and sample s, or set to 1 if at least one protein coding mutation was present for protein p and sample s. This matrix represented the presence of protein coding mutations in each protein coding gene in each sample.

2.2.2 Non-Metric Multi Dimensional Scaling

2.2.2.1 Creating the dissimilarity Matrix

Non-metric Multi Dimensional scaling (nmMDS) requires a pairwise sample dissimilarity matrix to be created from the binary mutation matrix. For each dataset I calculated the similarity between samples based upon the simple matching coefficient: the proportion of binary features in which two samples agree. I then subtracted each element of the simple matching similarity matrix from 1 to give the dissimilarity matrix required for nmMDS analysis.

2.2.2.2 nmMDS procedure

nmMDS uses a pairwise sample distance, or dissimilarity, matrix as its input to create a representation of the pairwise distance between all samples in as many dimensions as required by the researcher (up to the number of features used in the analysis). By representing pairwise sample distances in two dimensions I can represent pairwise similarities on a plot. A small distance between two points on the nmMDS plot indicated a high similarity between two samples. A plot of the two nmMDS dimensions resulting from a two dimensional nmMDS

procedure can be thought of as a close to optimal two-dimensional representation of the pairwise similarity between samples. In comparison to a plot of principal components, which are linear combinations of features in the data, nmMDS applies a mapping of samples to k-dimensional space, where k is defined by the user. This means two-dimensional nmMDS plots, where $k = 2$, can represent pairwise differences between samples based upon all features, which is not possible using two-dimensional plots of pairs of principal components. I conducted nmMDS using the ISOMDS function from the R "MASS" package (Venables & Ripley, 2003).

For all three datasets (TCGA June 2012, TCGA December 2012, and Pan-Cancer) I created nmMDS plots and coloured the points (samples) according to the sequencer analysis pipeline, and the micro-satellite instability status.

2.2.3 Identifying features correlated with the first principal component

Principal Component Analysis (PCA) was conducted on the binary mutation matrix for each of the three datasets. I used correlation analysis to identify whether the sequence analysis pipeline, age, gender, cancer type, and silent mutation frequency were correlated with the first principal component (PC1). If there was a technical effect present in the data, this analysis would identify that effect as a significant and large correlation with PC1.

2.2.4 Over-representation analyses to discover potential study design problems

I used the Fisher's Exact test, and Chi-square analysis to identify class imbalances across pairs of categorical features. I tested to see whether sequence technology was imbalanced across cancer anatomical site, cancer type, gender, age, and micro-satellite instability status.

The analysis scripts used in this chapter can be found at https://github.com/SutherlandRuss/RS_PhD_scripts.

2.3 Results

The Cancer Genome Atlas December 2012 data contained clinical and exome sequence data for 345 samples. The TCGA June 2012 and Pan-Cancer datasets contained the same 224 samples. There was an intersection of seventy-nine samples between the 224 Pan-Cancer and TCGA June 2012 data, and TCGA December 2012 data. The TCGA December 2012 hypermutated samples were not present in the Pan-Cancer dataset.

There was an intersection of 79 samples between the TCGA June 2012 and December 2012 datasets. There was a significant difference in the ranked number of silent mutations per individual ($Z = 45,\ p = 0.01$) between these two datasets. However, the median number of silent mutations was unchanged (Mdn= 17.0). The ranked number of mutations per individual changed across the two TCGA 2012 datasets ($Z = 253,\ p = 2.39 \times 10^{-5}$). Although, the median remained unchanged (Mdn=64).

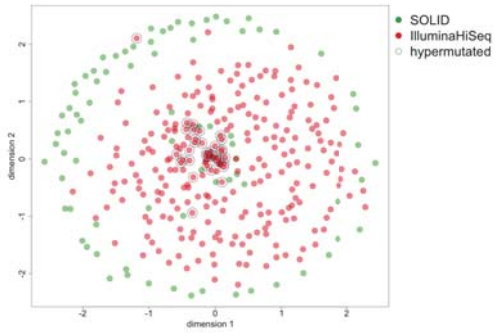
2.3.1 non-metric Multi-Dimensional Scaling

Figures 2.3.1a and 2.3.2a on page 61, shows a sequencing technology effect in TCGA December 2012 and TCGA June 2012 data respectively, which was not present in the Pan-Cancer data (Figure 2.3.3a). In both the TCGA December 2012, and TCGA June 2012 data, the majority of the hypermutated samples formed a cluster in the centre of the nmMDS plot (Figure 2.3.1a, and Figure 2.3.2a respectively). In TCGA December 2012 data the 55 microsatellite instability (MSI-H) samples were concentrated in the centre of the nmMDS plot (Figure 2.3.1b), as they were in the TCGA June 2012 data (Figure 2.3.2b). MSI-H status was associated with hypermutation ($p = 2.2 \times 10^{-16}$) in the TCGA June 2012 data. None of the TCGA December 2012 SOLiD sequenced MSI-H samples were hypermutated by my definition, although they seemed to be clustered at the centre of the samples (Figure 2.3.1b). The Pan-Cancer dataset did not include any of the 55 TCGA December 2012 hypermutated samples. As part of the Pan-Cancer quality control process hypermutated samples which carried more than 500 mutations were removed from further analysis. The hypermutated samples were expected to harbour many passenger mutations,

which do not contribute to tumour differentiation or progression (Kandoth et al., 2013a). These samples were removed by Kandoth et al. (2013a) to reduce noise when identifying significantly mutated genes and grouping samples based on common patterns of mutation. The Pan-Cancer data included nine MSI-H samples which when labelled according to the TCGA June 2012 metadata appeared to be clustered at the centre of the samples (Figure 2.3.3b).

Figure 2.3.1: TCGA December 2012 colorectal nmMDS plots

(a) Sequencing technology effect in TCGA December 2012 dataset using non-metric MDS



(b) Microsatellite instability-high central cluster in TCGA December 2012 dataset using non-metric MDS

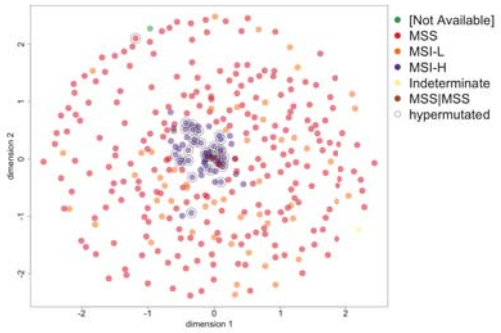
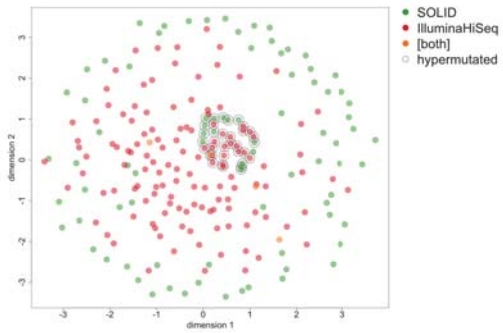


Figure 2.3.2: TCGA June 2012 colorectal nmMDS plots

(a) Sequencing technology effect in TCGA June 2012 dataset using non-metric MDS



(b) Microsatellite instability-high samples central cluster in TCGA June 2012 dataset using non-metric MDS

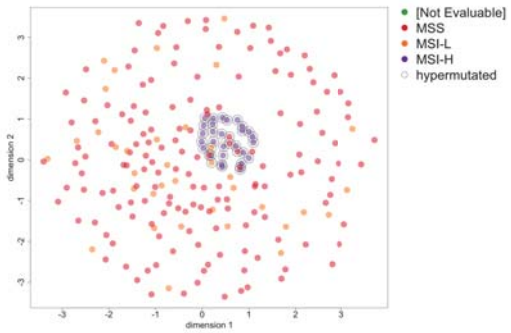
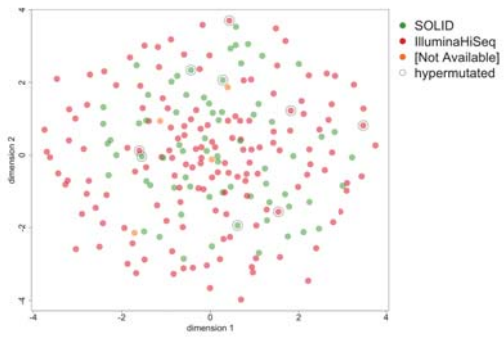
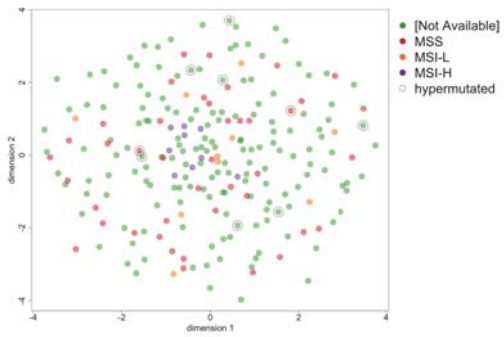


Figure 2.3.3: Pan-Cancer colorectal nmMDS plots

(a) Sequencing technology effect not seen in the Pan Cancer dataset using non-metric MDS



(b) Microsatellite instability-high samples central cluster in Pan-Cancer dataset using non-metric MDS



2.3.2 Metadata correlations with the first principal component.

In each of the datasets; TCGA December 2012, TCGA June 2012, and Pan-Cancer 2013; the first principal component of the variation in the binary mutation matrix was highly correlated with the silent mutation rate (Table 2.3.1). In each dataset there was also a small, but significant correlation between sequencing technology and the first principal component (Table 2.3.1).

Table 2.3.1: First principal component correlations with metadata in each data set.

	variance explained	N samples	silent mutation rate		sequencing technology	
			r_s	p-value		p-value
TCGA December 2012	7.08% (34.67)	345	0.96	2.2×10^{-16}	W = 17974	2.2×10^{-16}
TCGA June 2012	17.52% (51.40)	224	0.87	2.2×10^{-16}	X^2 (2) = 52.91	3.24×10^{-12}
Pan-Cancer analysis	16.68% (50.23)	224	0.86	2.2×10^{-16}	X^2 (2) = 9.46	0.01
	age		gender		cancer type	
TCGA December 2012	r_s	p-value	W	p-value	W	p-value
TCGA June 2012	0.03	0.77	15536	0.36	15521	1.19×10
Pan-Cancer analysis	0.19	4.28×10^{-3}	6968	0.14	6780	1.39×10^{-3}

For each dataset, relationships between; silent mutation rate, sequencing technology, age, gender, and cancer type; and PC1 were calculated. Wilcoxon's rank sum test was used to compare PC1 scores across cancer type and gender. Spearman's rho correlation was used to assess the relationship between PC1 and: silent mutation rate, and age. The kruskal wallis test was used to find differences in PC1 scores across sequencing technology for TCGA June 2012 and Pan-Cancer analysis data where there were three sample classes (SOLiD, Illumina, and both or not available). Wilcoxon's rank sum was used to find differences in PC1 scores across sequencing data for TCGA December 2012 data where there were two sequencing classes (SOLiD, and Illumina). In each data set the silent mutation rate is most highly correlated with the first principal component.

Across each of the three datasets PC1 scores differed across sequencing technologies, and cancer type (Table (2.3.1)). In The TCGA June 2012 and Pan-Cancer data there was a weak, but significant relationship between age and PC1 score.

2.3.3 Class imbalance analysis

The proportions of colon and rectum cancers processed using SOLiD, or Illumina technology were not balanced in the TCGA June 2012 and Pan-Cancer datasets (χ^2 (2, $N = 224$) = 12.60, $p = 1.84 \times 10^{-3}$). The anatomical site, the section of colon or rectum from which the tumour sample was collected, was not balanced across the sequencing technologies (χ^2 (8, $N = 224$) = 15.37, $p = 0.05$). However, age, gender and stage were balanced across the sequencers. In the TCGA December 2012 data the samples sequenced using Illumina technology were enriched for colon cancer samples ($p = 0.01$). The anatomical site of the cancers was imbalanced across sequencing technology (χ^2 (10, $N = 345$) = 29.87, $p = 9.01 \times 10^{-4}$), but age, gender, and stage were not.

2.4 Discussion

According to the nmMDS plots (Figures 2.3.1a and 2.3.2a) there was a sequencing technology effect in the TCGA December 2012, and TCGA June 2012 data, which was reduced in the Pan-Cancer 2013 data (Figure (2.3.3a)). The main driver of variation in the samples was the mutation rate, not a technical effect. PC1 and sequencer type were correlated in each of the three datasets, but this may be confounded by the skewed distribution of cancer types across the sequence technologies.

I expected to see identical number of mutations for the 79 samples that intersected between the TCGA June 2012 and TCGA December 2012 datasets. However, the number of mutations per sample changed across the two datasets, and indicated undocumented changes to the analysis pipeline.

The imbalance of colon and rectum cancers across sequencing technology may have

arisen because the study used an opportunity sample, rather than defining quotas at the study outset. Hypermutation was an important feature in the TCGA December 2012 and TCGA June datasets, and it was highly associated with the MSI-H phenotype A classifier could be developed to classify samples as MSI-H or non-MSI-H using exome sequence data and our sample similarity measure. However, there are already existing tools to predict MSI status using exome sequence data without using the established Bethesda marker panel test (RepeatSeq (Highnam et al., 2013), lobster (Gymrek et al., 2012)).

2.5 Conclusion

In each of the three datasets (TCGA December 2012, TCGA June 2012, and Pan-Cancer) the feature most highly correlated with PC1 of the binary mutation matrix was the frequency of mutations. Based on nmMDS plots there was a sequencing technology effect in the TCGA December 2012 dataset, and June 2012 TCGA data, which was reduced in the Pan-Cancer data set.

Future large scale sequencing studies should be aware of potential technical effects, and take appropriate precautions during the study design phase. Features which are known to exhibit variation across classes, should, as much as is possible, be balanced appropriately across different experimental conditions. I advise other researchers to use a single sequencing technology and pipeline to process all samples in a study.

Based on this exploratory investigation in to the TCGA and Pan-Cancer colorectal cancer data I decided that all further TCGA analyses in this thesis would use the Pan-Cancer data set.

Chapter 3

Predicting cancer grade and stage across three types of adenocarcinoma using exome sequence data

3.1 Introduction

The stage and grade of a tumour are deemed the strongest indicators of cancer prognosis (Engers, 2007). Cancer staging is conducted using the World Health Organisation (WHO) TNM system (Edge & Compton, 2010). It measures three standardised features of the cancer: the size of the primary tumour (T), the presence or absence of tumours in regional lymph nodes (N) and the presence or absence of distant metastases (M). There are four broad cancer stages, ranging from stage I to stage IV. The tumour size and cancer spread increases along with the cancer stage category.

Tumour grading is based solely on morphological criteria following histological assessment and, as such, is considerably more subjective than staging. Tumour grading measures the degree of differentiation of the neoplastic cells compared with the cells of the surrounding healthy tissue (Epstein, 2010). The degree of differentiation is a continuum from near normal to totally undifferentiated. Tumours with near normal differentiation are assigned a *low-grade* annotation (usually grade 1, or 2), and poorly differentiated tumours are assigned a *high-grade* annotation (usually grade 3, or 4). The cut-off points between the levels of differentiation are not well defined. Consequently, the definition for what is deemed well differentiated for one pathologist may well be different from another (Engers, 2007). Indeed, inter pathologist grading agreement is fair at best across multiple studies (Engers, 2007; Lang et al., 2005; Han et al., 2013; Scholten et al., 2004). In addition, multiple grading systems are used across and within cancer types. Prostate cancers use the Gleason grading system (Epstein, 2010), ovarian, and endometrial cancers use the FIGO grading system (Shepherd, 1989), and renal clear cell carcinomas use the Fuhrman grading system (Fuhrman, Susan A and Lasky, Larry C and Limas, 1982). A more standardised and objective measure may improve diagnostic accuracy and aid decision making regarding treatment. In order to achieve this, I must establish molecular correlates of cancer grade across multiple cancers.

The reduction in sequencing cost in the last decade has made possible the collection of large numbers of exome sequenced cancer samples. The Pan-Cancer Initiative (Weinstein et al., 2013) was established to enable the comparison of the first 12 Cancer Genome Atlas

cancer types across multiple omics data types. New cancer subtype classifications have resulted from the analysis of this dataset (Ciriello et al., 2013; Hofree et al., 2013). In this study I used data from the Pan-Cancer Initiative (Weinstein et al., 2013) to identify clinical and exome sequence-based features that are associated with cancer grade and stage across and within multiple types of adenocarcinomas, the most frequent epithelial subtype of cancer. Identification of exome-sequence based features that are correlated with cancer grade may help with the future development of a grading system that could be used across cancers, or at least help with our understanding of the genes involved in cancer grade.

There is limited literature on the sequence-based correlates of tumour grading within and across tumour tumour types in contrast to the well studied area of sequence-based correlates of tumour types. The existing literature use candidate gene approaches (Garcia-Dios et al., 2013). I could not find any previous work that attempted to predict cancer grade across multiple cancer types using clinical features and whole exome sequence derived features. Lee et al. (2013) created models to predict advanced clinical stage in TCGA colorectal cancer data. They used an elastic net (Zou & Hastie, 2005) approach to integrate mutation, gene expression, copy number alteration and methylation data and found that a set of 158 features corresponding to 143 genes gave the best model of association. However, they did not validate their models in an independent test set, nor did they attempt to identify the common correlates of cancer stage across multiple cancers. Greater understanding of the genes involved in cancer grade could help lead to the creation of an objective tumour grading system across cancers that would be useful to pathologists when performing tumour grading, and may be applicable to tumours which currently do not have grading systems such as adrenocortical tumours.

3.2 Methods

I used patient age (A; at initial onset), gender (G), cancer stage (S), tumour type (T), and features derived from the exome somatic mutation profile of each patient's primary tumour to predict whether the tumour was low-grade (1,2) or high-grade (3,4). Logistic regression and AIC (Akaike, 1974) backwards model selection were used to create the models in the presence or absence of relevant clinical data such as tumour staging. The analyses were repeated using stage as the outcome of interest.

Adenocarcinoma grade predictive models were built by combining data from the three tumour types (endometrial carcinoma, renal cell carcinoma and ovarian carcinoma) and then in each individual tumour type separately. A summary of the model building steps is provided in Figure 3.2.1 on page 72. Four 'baseline' models were built comprising the clinical features (A,G,S,T; Table 3.2.1). Models were re-run with the inclusion of exome-derived features, namely protein mutational information (P) and variant frequency features (V). By including the exome sequence derived features in addition to the clinical features I could measure any additional predictive power captured by the exome sequence derived features. For ovarian carcinoma and endometrial carcinoma models were built without inclusion of gender.

Table 3.2.1: Grade and stage classification model abbreviations

Across cancer models	**Model name**	**+ proteins**	**+ variant frequency**	**+ proteins & variant frequency**
age gender	AG	AGP	AGV	AGPV
age gender stage	AGS	AGSP	AGSV	AGSPV
age gender tumour type	AGT	AGTP	AGTV	AGTPV
age gender stage tumour type	AGST	AGSTP	AGSTV	AGSTPV
Within cancer models	**Model name**	**+ proteins**	**+ variant frequency**	**+ proteins & variant frequency**
age	WA	WAP	WAV	WAPV
age (gender)	WAG	WAGP	WAGV	WAGPV
age (gender) stage	WA(G)S	WA(G)SP	WA(G)SV	WA(G)SPV

A summary of the logistic regression models combining clinically derived features with exome derived mutation features. **Model name** = models composed of clinical features are baseline models. + **proteins** = baseline clinical and protein feature models. + **variant frequency** = baseline clinical and variant frequency feature models. + **proteins & variant frequency** = baseline clinical , protein and variant frequency models.For stage prediction models the stage covariate (S) in the above models was replaced with the grade covariate (Gr).

3.2.1 The data set

Mutation Annotation Format (MAF) files and clinical data files for each of the 12 Pan-Cancer project data sets (accession syn1710680) were downloaded from Sage Synapse on 3rd October 2013. For each individual exome, sequencing had been performed on paired normal and tumour tissue in order to identify the mutations specific to the tumour tissue (somatic mutations). I retained 970 samples for which exome-sequence data and age, gender, tumour stage and tumour grade data were available.

3.2.2 Data pre-processing

3.2.2.1 Phenotype definition: Tumour grade and stage grading dichotomisation

Eight tumour grading categories were used to define grade in the Pan-Cancer data (G1, G2, G3, G4, Grade 1, Grade 2, Grade 3, High-Grade). The tumours with a grading equivalent to grades 1 or 2 were assigned to the *low-grade* category. The *high-grade* tumours had grade classifications of G3, G4, Grade 3 and High-grade. In renal clear cell carcinoma studies, this approach yielded improved inter-pathologist grading agreement (Lang et al., 2005; Al-Aynati et al., 2003). The distribution of grade annotations across tumour type and stage (low/high) is shown in Table 3.6.4. Twenty two cancer stage categories were used to describe the cancer stage of the 970 tumour samples. I grouped the 22 classes into two classes, *low-stage* (stages I and II) and *high-stage* (stages III and IV). Low-stage samples represented the cancers limited to the organ of origin without evidence of metastasis, and high-stage samples represented locally advanced and metastatic cancers. The distribution of stage annotations across tumour type and grade (low/high) is shown in Table 3.6.3.

3.2.2.2 Predictors: Creating the protein binary mutation matrix

I removed 'silent' mutations, which did not change the amino acid sequence in the translated protein, from the MAF file. I used the union of the functionally mutated proteins across

all samples to generate a matrix M of proteins P by samples S. Each element in $M[p, s]$, was set to 0 if there was no protein coding mutation for protein p, and sample s, or set to 1 if at least one protein coding mutation was present for protein p, and sample s. This matrix represented the presence of any protein coding mutations in each protein coding gene in each sample. Similar approaches have been taken in other studies (Hofree et al., 2013; Leiserson et al., 2013a).

3.2.2.3 Predictors: Variant frequency features

I recorded the frequency of each of the six transition and transversion SNV substitutions across the exome to test for their ability to discriminate low-grade from high-grade tumours. Transition and Transversion frequencies have been shown to discriminate between cancer types (Kandoth et al., 2013a; Lawrence et al., 2013). I also recorded, and tested the discriminatory potential of the frequency of ten types of sequence variant; frame shift deletions, frame shift insertions, in-frame deletions, in-frame insertions, missense, nonsense, RNA, silent, and splice site mutations. The variant frequency features were not normally distributed across samples (Figure 3.6.1) therefore, I tested whether median variant frequencies varied across grade using the Wilcoxon rank sum test, and across stage using the Kruskal-Wallis test.

3.2.2.4 Defining the training set and test set

For each of the analyses, I randomly assigned samples without replacement to a 2/3 training set and a 1/3 test set. I balanced the high-grade and low-grade outcomes in the training sets to prevent a binary classifier from favouring the majority class (Barandela & Sã, 2003). I down-sampled the majority class to be of equal size to the minority class. Any surplus majority class training samples were assigned to the test set.

3.2.3 Mutated protein and variant frequency feature selection

I reduced the number of proteins used as predictors in the training sets using a simple heuristic; a protein feature was retained if it carried at least one protein coding mutation in

more than 5% of training samples in any one of the tumour types, similar to the approach used by Kandoth et al. (2013a). The number of proteins retained for each analysis is shown in Table 3.2.2. I then built single exome feature logistic regression models for each of the remaining proteins and each variant frequency feature separately including clinical features age, gender, stage, grade and tumour type (for the cross tumour models) as covariates.

Table 3.2.2: Number of proteins retained after frequency filter

	grade classification			
	All 3 cancers	**renal cell**	**ovarian carcinoma**	**endometrial carcinoma**
training set size	474	236	38	130
N genes passing filters	562	12	16	1788
	stage classification			
	All 3 cancers	**renal cell**	**ovarian carcinoma**	**endometrial carcinoma**
training set size	580	238	18	84
N genes passing filters	919	14	47	2360

Across the adenocarcinomas and within each adenocarcinoma, the size of the training set used to build each logistic regression models are shown above. The **N genes passing filters** shows the number of genes which carry functional mutations in at least 5% of the training set.

3.2.4 Final model building

All features passing a threshold for significance in the single exome feature logistic regression (FDR<0.1 (Benjamini & Hochberg, 1995)) were taken forward for a range of multi-exome feature logistic analyses. I used stepwise backward model selection (using Akaike's information criterion (Akaike, 1974)) to refine the multi-exome feature models further for each of the sets of features. Stepwise backwards model selection removed each feature from the logistic regression model and re-computed Akaike's Information Criterion (AIC), if the AIC was lower in the new model, the new model was adopted as the 'new best' model. This process continued until the AIC could not be lowered by removing any features from the 'new best' model. At that point the model was considered optimal. It must be remembered that here 'new best' refers to a local optima and not the optimal global solution to the regression problem.

3.2.4.1 Model performance and comparisons

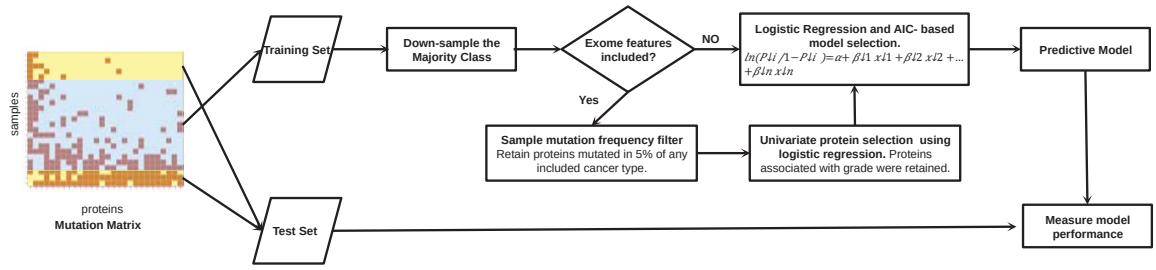
I used model sensitivity, specificity, accuracy, and the area under the curve (AUC) to measure and compare the performance of each classification model along with model sensitivity, specificity and accuracy. I tested whether any improvement in classification performance was achieved by including exome derived sequence features over clinically derived features alone using the roc.test function within the pROC R package (Robin et al., 2011). I used the bootstrap permutation method originally described in Hanley & McNeil (1983). A p-value below 0.05 indicated an improvement in classification performance by adding exome sequence derived information to the clinical variables in the 'baseline' models AGS, AGT, and AGST.

3.2.4.2 Tumour stage classification methods

I used the same model building process as described above (sections 2.1-2.4.2) to predict cancer stage, this time including tumour grade (low-grade / high-grade) as a predictor.

The analysis scripts used in this chapter can be found at https://github.com/SutherlandRuss/RS_PhD_scripts.

Figure 3.2.1: Grade and stage classification analysis pipeline



The analysis pipeline shows how I first split the samples in to a two thirds training set and one third test set. The majority class was downsampled to the minority class size. If the model was to include protein features then all proteins with mutations in $\geq$5% of the training set were retained. Individual proteins which were significantly associated with the outcome were retained for final multi-protein model building. If proteins were not to be included in the model, the 5% mutation frequency and logistic regression feature selection steps were skipped. The final multivariate model building step resulted in AIC refined model which was then tested using the test set.

3.3 Results

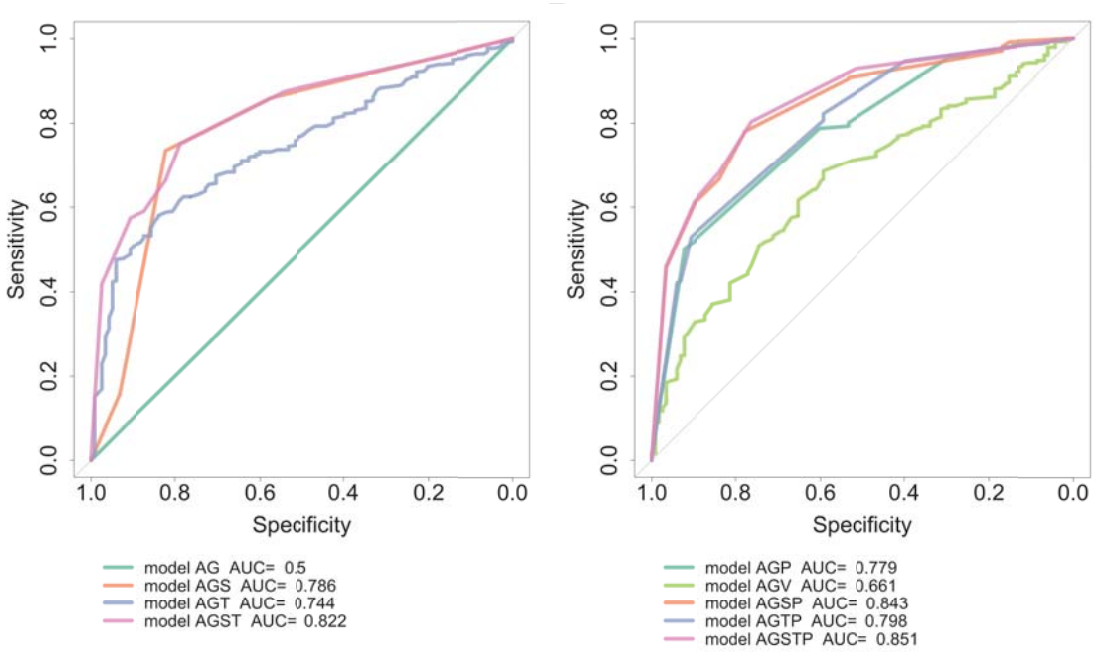
I used 970 samples (247 endometrial carcinoma, 416 renal cell carcinoma and 307 ovarian carcinoma) from the Pan-Cancer data set for which exome sequence data, age, gender, tumour stage and tumour grade information was available from the 3908 Pan-Cancer data set samples. Three hundred and fifty five samples were annotated as low-grade, and 615 as high-grade across the three tumour types (150 low, 97 high for endometrial carcinoma; 177 low, 239 high for renal cell carcinoma; 28 low, 279 high for ovarian carcinoma). There were 435 low-stage (stage 1 and stage 2) and 535 high-stage (stage 3 and stage 4) samples across the 3 tumour types (184 low, 63 high endometrial carcinoma; 237 low 179 high renal cell carcinoma; 14 low 290 high for ovarian carcinoma). The sample demographics and summary statistics for endometrial carcinoma, renal cell carcinoma and ovarian carcinoma combined and stratified according to cancer stage, high-grade / low-grade categorisation are shown in Table 3.6.1. The distribution of grade classes across cancer types and stages (low/high) is shown in Table 3.6.3 along with the grade class mapping. The distribution of stage classes across cancer types and grades (low/high) is shown in Table 3.6.4 along with stage class mapping. Across adenocarcinomas frame-shift deletions, frame-shift insertions, in-frame deletions and splice site mutations showed significant differences in frequency between low and high-grade tumours. Wilcoxon rank sum tests are available in Table 3.6.2. However, none were predictive of grade when adjusted for age, stage, tumour type and proteins in the regression models. For the endometrial cancers, C>G/G>C transversion mutation frequency was higher in the low-grade tumours in comparison to high-grade tumours after Bonferroni correction ($0.05/64$, $\alpha = 7.8 \times 10^{-4}$), $W(n_1 = 150, n_2 = 97) = 4918.5, p = 2{\times}10^{-5}$. None of the variant frequency features showed a significant difference across grade, or stage in the renal or ovarian carcinomas.

3.3.1 Cross-adenocarcinoma grade classification

When predicting grade across adenocarcinomas, the presence of at least one *TP53* coding mutation was predictive of high tumour grade ($OR = 7.62$, $CI[4.35, 13.64]$, $p = 2.91 \times 10^{-14}$)

when adjusting for patient age (at initial onset), gender ($OR = 2.89$, $CI[1.74, 4.84]$, $p = 4.76\times10^{-5}$), and stage ($OR = 4.89$, $CI[3.08, 7.84]$, $p = 2.61\times10^{-11}$) in the best performing model AGSTP. This model performed significantly better than the clinical features alone. ($D = 3.25$, $p = 6.0 \times 10^{-4}$). ROC curves are shown in Figure3.3.1. Across adenocarcinomas in model AGTP, *TP53* protein coding mutations were predictive of cancer grade ($OR = 7.40$, $CI[3.60, 15.73]$, $p = 8.84\times10^{-08}$) when adjusted for patient gender ($OR = 2.56$, $CI[1.38, 4.86]$, $p = 3.29\times10^{-03}$), and tumour type (ovarian tumour type [$OR = 2.46$, $CI\{0.95, 6.36\}$, $p = 0.06$], endometrial tumour type [$OR = 0.62$, $CI\{0.31, 1.25\}$, $p = 0.18$]). Model AGTP performed better than the AGT model ($D = 3.45$, $p = 3.0 \times 10^{-4}$). In the AGV model C>G/G>C transversions were predictive of cancer grade ($OR = 1.06$, $CI[1.03, 1.10$, $p = 8.21\times10^{-03}$]), as were frame shift deletions ($OR = 0.88$, $CI[0.83, 0.93]$, $p = 6.13\times10^{-06}$). When proteins were included with age, and gender (model AGP), the variant frequency features were replaced in the model by *TP53* ($OR = 12.24$, $CI[6.89, 22.46]$, $p = 2.0\times10^{-16}$), *ARID1A* ($OR = 2.13$, $CI[0.97, 4.70]$, $p = 0.06$) , *CTCF* ($OR = 0.36$, $CI[0.08, 1.18]$, $p = 0.13$), *CTNNB1* ($OR = 0.09$, $CI[0.01, 0.33]$, $p = 1.69\times10^{-03}$) and *PIK3R1* ($OR = 0.41$, $CI[0.15, 1.04]$, $p = 0.69$). I found differences when comparing the variant frequency features across cancer grade at the univariate level, but when included in the regression and adjusted for all covariates they were not predictive of cancer grade (Table3.6.2). Cross-cancer grade classification statistics can be found in Table 3.3.1.

Figure 3.3.1: Across cancer grade classification ROC curves.



A= age, **G**= gender, **S**= Stage, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency. Across cancer models to predict cancer grade based upon clinical features perform well with model AGST performing best (AUC = 0.822). When the clinical models are supplemented with protein information classification performance improves slightly and the best performing model is AGSTP (AUC = 0.851).

Table 3.3.1: Across cancer grade classification statistics

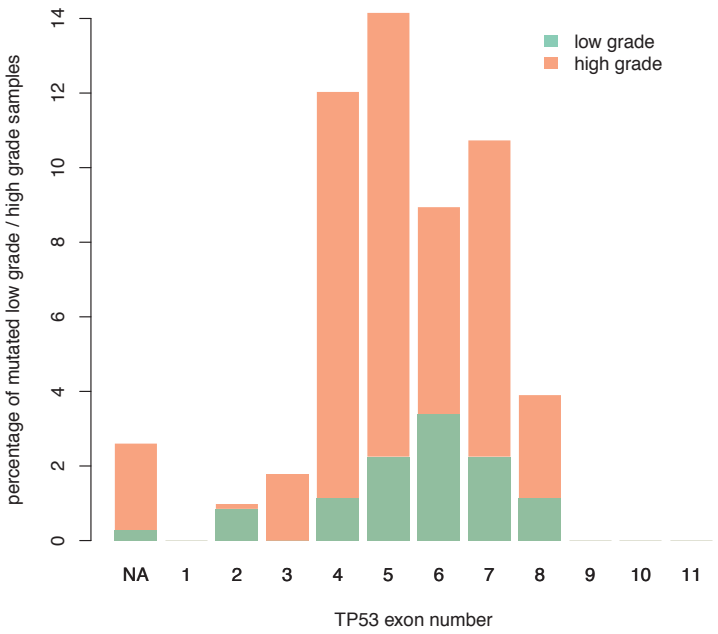
	Sensitivity	Specificity	Accuracy	Positive predictive value (PPV)	Area Under the Curve (AUC)
model AG	0.50	0.50	0.50	NA	0.50
model AGP	0.53	0.90	0.62	0.94	0.78
model AGV	0.68	0.61	0.66	0.85	0.66
model AGPV	0.53	0.90	0.62	0.94	0.78
model AGS	0.74	0.82	0.76	0.93	0.79
model AGSP	0.78	0.78	0.78	0.92	0.84
model AGSV	0.74	0.82	0.76	0.76	0.79
model AGSPV	0.78	0.78	0.78	0.92	0.84
model AGT	0.59	0.81	0.64	0.91	0.74
model AGTP	0.53	0.91	0.62	0.95	0.80
model AGTV	0.59	0.81	0.64	0.91	0.74
model AGTPV	0.53	0.91	0.62	0.95	0.80
model AGST	0.66	0.82	0.70	0.92	0.82
model AGSTP	0.68	0.84	0.72	0.93	0.85
model AGSTV	0.66	0.82	0.70	0.92	0.82
model AGSTPV	0.68	0.84	0.72	0.93	0.85

A= age, **G**= gender, **S**= Stage, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency.

Model performance metrics for all models created to predict cancer grade across endometrial carcinoma, ovarian carcinoma and renal cell carcinoma. Classification sensitivity was highest for those models that included cancer stage. Model AGTP achieved a high-grade classification specificity of 0.907 in the test set, indicating that any unseen tumour classified as high-grade is likely to be a high-grade tumour. For models with identical classification statistics, such as AGSTP and AGSTPV, the more complex model (AGSTPV) has been refined to the simpler model (AGSTP) though the AIC model selection process.

I investigated the positions of *TP53* mutations by mapping mutations to the 11 *TP53* exons (Figure3.3.2). At each TP53 exon a larger percentage of high-grade samples carried mutations than low-grade samples. Clinical sequencing of *TP53* has typically focused on exons 5-8. In addition to mutations in exons 5-8, 31.6% of *TP53* mutations in high-grade samples mapped outside exons 5-8. Among the low-grade samples 20% of *TP53* mutations mapped outside of exons 5-8. The vast majority of samples carried a single protein coding mutation in *TP53*. There were 16 mutations among high grade samples and a single mutation among low-grade samples that did not map to the canonical TP53 exons.

Figure 3.3.2: Percentage of low-grade and high-grade samples carrying protein coding mutations in each *TP53* exon.



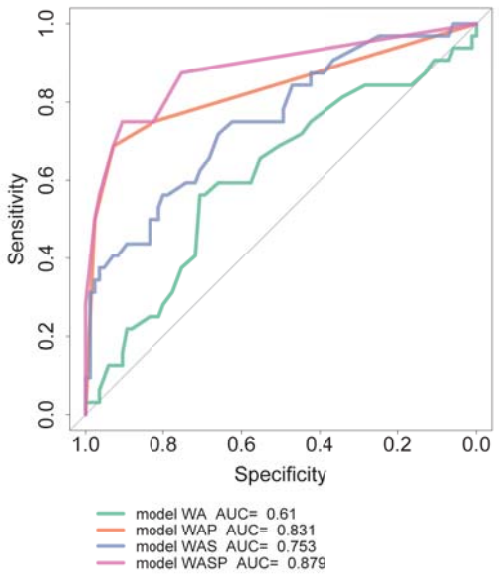
The *TP53* gene has eleven canonical exons based upon the ensembl release 76 (Flicek et al., 2014) transcript ENST00000269305 represented on the x-axis. In high-grade samples *TP53* mutations are more prevalent than in low-grade samples across most exons. *TP53* mutations are found outside of exons 5-8 which are typically sequenced in the clinic. 'NA' mutation frequency represent those mutations that did not map to any *TP53* exon.

3.3.1.1 Grade classification within adenocarcinomas

When predicting endometrial carcinoma grade the best performing model included age, stage, and protein features (model WASP) (Table 3.6.9 on page 94) with AUC of 0.879 (Figure3.3.3). *TP53* functional mutations were associated with high cancer grade (OR = 11.78, CI[3.91, 44.28], $P = 4.64 \times 10^{-05}$) when adjusted for cancer stage (OR = 2.06, CI[0.77, 5.57], $p = 0.15$) (Table 3.6.7 on page 92). *PTEN* functional mutations predicted low cancer grade ($OR = 0.38$, CI[0.15, 0.94], $p = 0.04$). The WAP model outperformed the baseline model WA ($D = 3.42$, $p = 3.0 \times 10^{-4}$), as did model WAGSP ($D = 2.5$, $p = 0.01$) (Figure 3.3.3 on page 79, and Table 3.6.9 on page 94). The WASP model outperformed the WAP model ($D = 2.17$, $p = 0.02$).

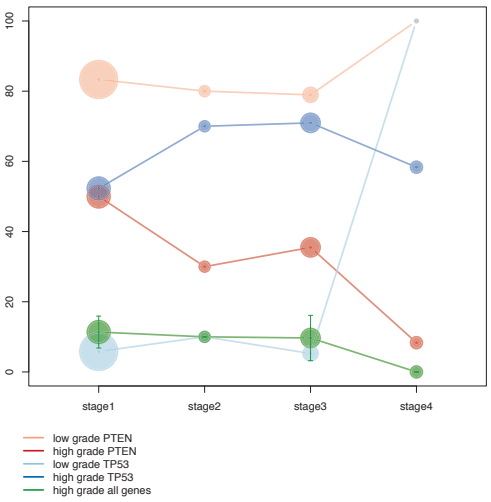
Histological type was not included in the logistic analysis because of the distribution of grade across histological type. After I removed 17 samples of indeterminate histological type with "serous and endometrioid features", there remained 150 low-grade and 49 high-grade endometrioid adenocarcinomas. For serous tumours there were 44 high-grade tumours and no low-grade tumours. I could not reliably measure the effect of *TP53* and *PTEN* mutations. I found that samples with functionally mutated copies of *TP53*, and *PTEN* were imbalanced across histological type using Fisher's exact test (*TP53* $p = 3.8 \times 10^{-8}$, *PTEN* $p = 5.55 \times 10^{-10}$). I investigated whether *TP53* and *PTEN* sample mutation frequency changed across stage for high-grade and low-grade tumours (Figure3.3.4), and found a negative association between *PTEN* functional mutation and increasing cancer stage in high-grade endometrial carcinomas ($OR = 0.61$, $CI[0.41, 0.90]$, $P = 0.02$). A trend that was reflected across other genes in the high-grade samples (Figure 3.3.4). For the ovarian and renal adenocarcinomas, neither proteins, nor variant frequency features were included in cancer grade prediction models after AIC backwards model selection (Figures 3.6.3a, and 3.6.4a on page 97; Table 3.6.7 on page 92, Table 3.6.10 on page 95, and Table 3.6.11 on page 98)

Figure 3.3.3: Endometrial carcinoma grade classification ROC curves



W= within cancers, **A**= age **S**= Stage, **P**= proteins. Model WAP performs comparably to the WASP model, despite not including stage information.

Figure 3.3.4: Endometrial carcinoma sample mutation frequency.



Circle sizes indicate sample sizes. Larger circles indicate a larger sample for that particular group. The low-grade stage four group was composed of a single sample, which carried functional mutations in both *TP53* and *PTEN*.

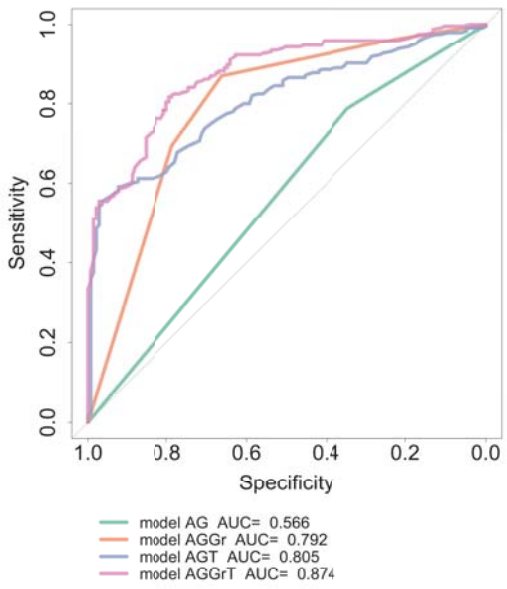
3.3.2 Stage classification

Across adenocarcinomas the age, gender, grade, tumour type and proteins model (model AGGrTP) performed with the highest AUC of 0.874 (Figure 3.3.5b). The significantly contributing features were tumour grade ($OR = 4.50$, $CI[2.89, 7.11]$, $p = 5.14\times10^{-11}$), ovarian tumour type ($OR = 20.50$, $CI[10.6, 43.87]$, $p = 2.0\times10^{-16}$) and endometrial tumour type ($OR = 5.10$, $CI[0.30, 0.84]$, $p = 8.91\times10^{-03}$). Model AGGrP performed with an AUC of 0.847 (Figure 3.3.5b). The predictive features were high-grade ($OR = 4.72$, $CI\,[3.00, 7.52]$, $p = 3.33\times10^{-11}$), male gender ($OR = 0.59$, $CI\,[0.35, 0.98]$, $P = 0.04$), *TP53* ($OR = 5.40$, $CI\,[3.12, 9.48]$, $p = 2.67\times10^{-09}$), *FNDC1* ($OR = 0.16$, $CI\,[0.03, 0.69]$, $p = 0.02$), *PIK3CA* ($OR = 0.21$, $CI\,[0.10, 0.42]$, $p = 1.88\times10^{-05}$) and *PIK3R1* ($OR =$

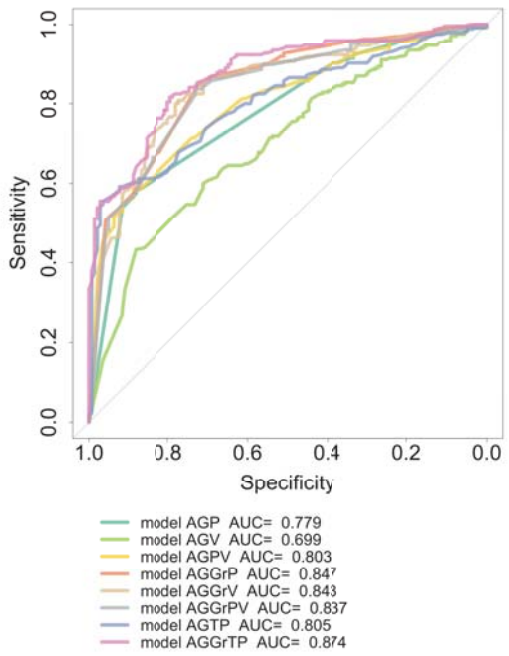
0.24, $CI\,[0.09, 0.57]$, $p = 2.06\times10^{-03}$). Model AGGrV achieved an AUC of 0.843 (Figure 3.3.5b) and the contributing features were; high-grade ($OR = 8.30$, $CI\,[5.53, 12.68]$, $p = 2.0\times10^{-16}$), male gender ($OR = 0.39$, $CI\,[0.25, 0.60]$, $p = 1.59\times10^{-05}$) and frame-shift deletion frequency ($OR = 0.92$, $CI\,[0.85, 0.98]$, $p = 0.01$). When proteins were included instead of cancer grade, model AGPV performed with an AUC of 0.803 (Figure 3.3.5b) and the predictive features were; *TP53* ($OR = 8.94$, $CI\,[5.71, 14.36]$, $p = 2.0\times10^{-16}$), *PIK3CA* ($OR = 0.24$, $CI\,[0.12, 0.45]$, $p = 1.72\times10^{-03}$), *and PIK3R1* ($OR = 0.24$, $CI\,[0.09, 0.56]$, $pp = 1.72\times10^{-03}$).

Figure 3.3.5: Stage classification across cancers ROC curves

(a) clinical feature baseline models



(b) clinical feature plus exome information models



A= age, **G**= gender, **Gr**= grade, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency. Across cancers the best performing model was AGGrT (AUC = 0.874). No additional predictive power was gained by including protein information in model AGGrTP.

Table 3.3.2: Across cancer stage classification statistics

	Sensitivity	Specificity	Accuracy	Positive predictive value (PPV)	Area Under the Curve (AUC)
model AG	0.79	0.35	0.62	0.67	0.57
model AGP	0.58	0.88	0.70	0.89	0.78
model AGV	0.71	0.54	0.65	0.72	0.70
model AGPV	0.58	0.88	0.69	0.89	0.80
model AGGr	0.87	0.66	0.79	0.81	0.79
model AGGrP	0.85	0.72	0.80	0.84	0.85
model AGGrV	0.82	0.75	0.80	0.85	0.84
model AGGrPV	0.82	0.73	0.79	0.84	0.84
model AGT	0.54	0.97	0.70	0.97	0.81
model AGTP	0.54	0.97	0.70	0.97	0.81
model AGTV	0.54	0.97	0.70	0.97	0.81
model AGTPV	0.54	0.97	0.70	0.97	0.81
model AGGrT	0.73	0.84	0.77	0.88	0.87
model AGGrTP	0.73	0.84	0.77	0.88	0.87
model AGGrTV	0.73	0.84	0.77	0.88	0.87
model AGGrTPV	0.73	0.84	0.77	0.88	0.87

A= age, **G**= gender, **Gr**= grade, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency. Model performance metrics for all models created to predict cancer stage across endometrial carcinoma, ovarian carcinoma and renal cell carcinoma. Classification performance as measured by AUC was highest for the age, gender, grade and tumour type model at 0.874. The AGPV model showed high-grade classification specificity of 0.874. For any unseen sample I would be confident that a high-grade classification was correct. For models with identical classification statistics, such as AGGrT and AGGrTPV, the more complex model (AGGrTPV) has been refined to the simpler model (AGGrT) though the AIC model selection process.

3.3.2.1 Stage classification within adenocarcinomas

No exome sequence features were predictive of cancer stage when adenocarcinoma types were analysed individually. In endometrial carcinoma age, and grade (model WAGr) (Table 3.6.9 on page 94, and Table 3.6.8 on page 93) was the best performing model, and had an AUC of 0.625 (Figure 3.6.2 on page 96). The WAGr model was also the best performer in ovarian carcinoma (Table 3.6.11 on page 98, and Table 3.6.8 on page 93) with an AUC of 0.661 (Figure 3.6.3b on page 96). The age, gender, and grade model (model WAGGr) (Table 3.6.10 on page 95, and Table 3.6.8 on page 93) performed best in renal carcinoma with an AUC of 0.716 (Figure 3.6.4b on page 97).

The complete set of cross cancer grade and stage prediction model summaries are shown in Tables 3.6.7 on page 92 and 3.6.8 on page 93 respectively. Cross cancer grade and stage model classification statistics are shown in Table 3.3.1 on page 76 and Table 3.3.2.

3.4 Discussion

This is the first study to attempt to use whole exome sequence derived features to identify the mutated genes associated with tumour grade across adenocarcinomas. I show potential for a model combining age, gender, tumour type and a protein mutation feature in the absence of tumour stage. When adding the protein feature (mutation of TP53) to the model, high specificity was achieved (0.907). The predictive accuracy of model AGSTP including *TP53* mutation status outperformed model AGST which was composed of clinical features.

This confirms the previously known association between *TP53* mutations and cancer grade in ovarian carcinoma and endometrial carcinoma. Across three types of adenocarcinoma (endometrial, kidney, and ovarian) I have shown that the presence of one or more protein coding mutations in *TP53* is associated with of high-grade cancer when adjusted for age, stage, gender and tumour type.

TP53 functional mutations were associated with high-grade status in endometrial carcinoma, and *PTEN* functional mutations were associated with low-grade cancer. However, the serous and endometrioid histological subtypes may confound this result: I could not adjust for histological type in the regression models because all except one serous endometrial carcinoma tumour was high-grade, leading to a breakdown of the logistic regression model. In endometrial carcinoma, the high-grade serous tumour type is characterised by near ubiquitous mutation of *TP53*, whereas mutation of *PTEN* is indicative of the endometrioid tumour type, these mutations being almost mutually exclusive (Kandoth et al., 2013a). However, *TP53* mutation incidence is greater in high-grade endometrioid tumours than in low-grade endometrioid tumours (Lax et al., 2000). The trend for *PTEN* mutation frequency to decrease across cancer stage when stratified by grade (Figure 3.3.4) may reflect an association between *PTEN* mutations and micro-satellite instability-high tumours, which are typically low-stage (Samarnthai et al., 2010; Kandoth et al., 2013b).

I discovered a set of proteins which when mutated were associated with high cancer stage, where low-stage indicated at worst a locally advanced tumour, and high-stage indicated tumour extension beyond the organ of origin. Using model AGPV, the sensitivity was 0.584,

but the high specificity of 0.876 shows that using a model comprising frame-shift deletions, *TP53, ARID1A, PIK3CA* and *PIK3R1* it may be possible to predict cancer stage using a simple statistical model that does not require surgery or a Magnetic Resonance Imaging (MRI) scan. Perhaps, further developed, this model could be used to prioritise patients with clinical feature and exome feature profiles that do not give a clear stage prediction for more extensive stage characterisation. *TP53* mutations are known to be associated with high cancer stage. *PIK3CA* and *PIK3R1* mutations predicted low tumour stage, which is consistent with the findings of showing co-occurrence of mutation in these genes if *PTEN* was also mutated (Kandoth et al., 2013b).

Even though I am the first to take a whole exome approach to identify mutated genes associated with cancer grade across adenocarcinomas I found no additional genes associated with cancer grade. I confirmed the association between *TP53* mutation and high-grade status in endometrial carcinoma, and the association of *PTEN* mutation with endometrioid histological type found in a candidate gene analysis (Garcia-Dios et al., 2013).

There was no pathologist ID attached to each tumour's grading in TCGA Pan-Cancer data, putting a limitation on this analysis. Some variation and noise should be expected from this fact. However, I expect that the most prognostic mutated genes should be common across pathologists. Up until recently, the original tissue histology reports were available, and it may have been possible to create a pathologist ID for the Pan-Cancer dataset based on the histology reports. However, in recent months the pathology report clinician signatures have been censored, and it is not possible to accurately assign each pathology report to a pathologist. For large scale cancer sequencing studies, such as Genomics England, it would be beneficial to record the pathologist ID so that the variation among pathologists can be modeled. It is entirely possible in this study that issues with inter pathologist grading reliability contributed to heterogeneity in the high-grade and low-grade classes. This may explain why *TP53* was the only protein I found to be associated with cancer grade. I attempted to account for this issue by reducing the number of grading classes from three and four class systems to a two class system which has been shown to improve inter pathologist

reliability (Scholten et al., 2004), without a reduction in prognostic power (Hong et al., 2011).

3.5 Conclusion

There is potential for an across cancer grading system using exome sequence data and clinical features. Using age, gender, tumour type and exome sequence protein features (model AGTP) to predict cancer grade performs almost as well as using age, gender, stage and tumour type information (model AGST). This system could be used to aid tumour grading when cancer stage information is not available, for example when an MRI has not yet been conducted.

Across adenocarcinomas *TP53* functional mutation was predictive of high-grade tumour status once adjusted for age, gender, stage and tumour type. Although across adenocarcinomas the exome sequence information did not add much predictive information above the clinical covariates in model AGST, tumour stage information is not typically available to pathologists when conducting tumour grading. There may also be potential to use the tumour stage prediction model AGPV to predict cancer stage across adenocarcinomas based on frame-shift deletion frequency and *TP53*, *ARID1A*, *PIK3CA* and *PIK3R1* features.

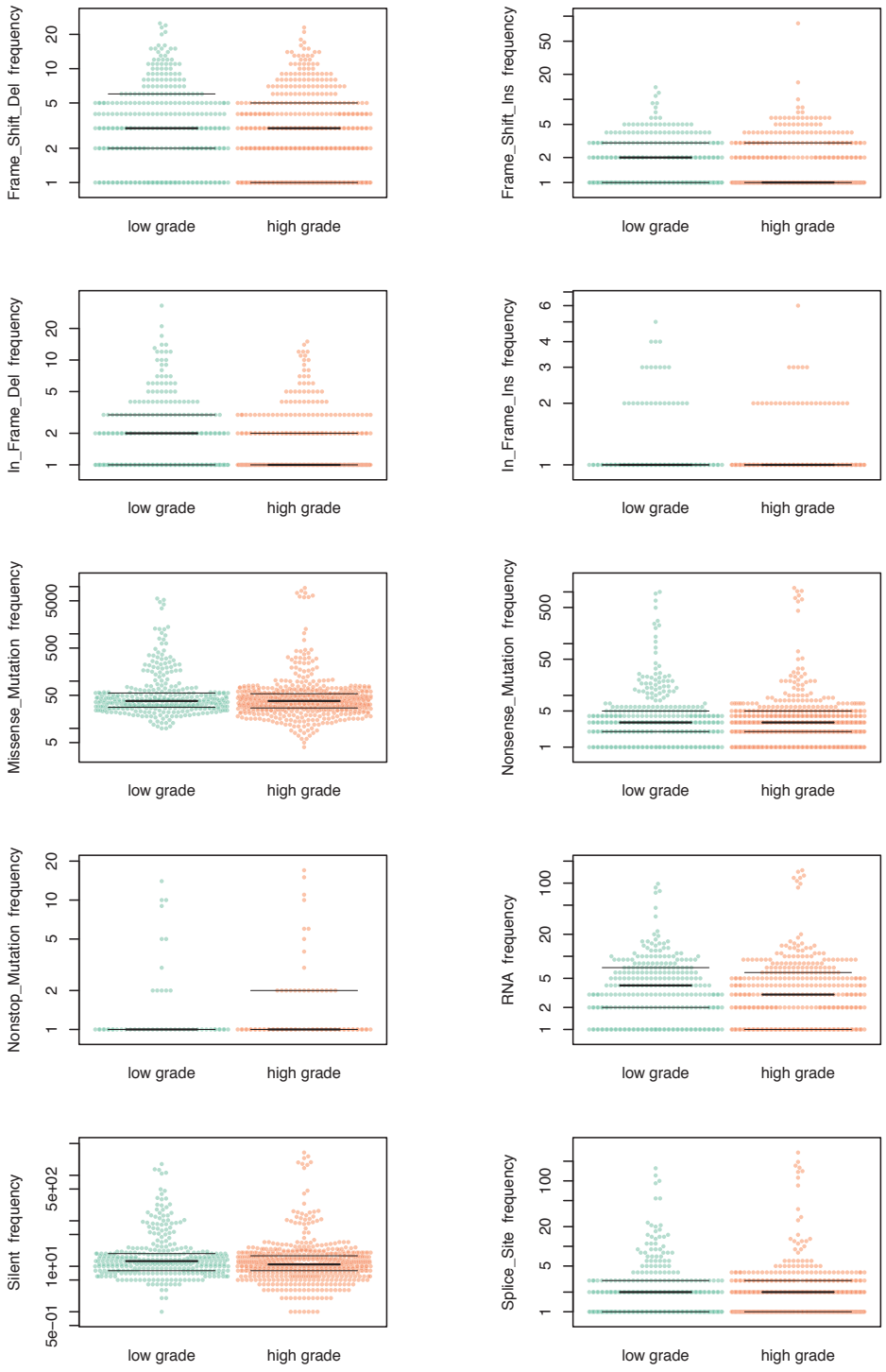
In endometrial carcinoma *TP53* and *PTEN* functional mutation was associated with cancer grade, although this was not independent of histological type. Across ovarian and renal carcinomas there were no proteins associated with cancer grade.

I would urge future large-scale cancer sequencing projects, such as Genomics England to record a pathologist ID variable along with tumour grade. This would allow scientists to create more robust tumour grade prediction models adjusted for pathologist effects. The reported genes and variant frequency features may reflect differing proportions of each cancer of high-grade and low-grade status included in the training set.

This study is being further developed. The TCGA pathology slides are being re-graded by a single expert pathologist (Dr Salvador Diaz-Cano) . These annotations will be used as the outcome in further regression models and will be free of inter-pathologist variation. Further work could involve a tumour grading validated by multiple pathologists, to create a *consensus grading* for each tumour type at which point machine learning methods would predict the *consensus grade*.

3.6 Supplementary materials

Figure 3.6.1: Variant types are non-normally distributed across cancer grade



Beeswarm scatterplots show that the frequency of mutations across samples is non-normally distributed for all variant types in low-grade and high-grade cancers. The median muteation frequency, and lower and upper quartile values are indicated thick and thin black lines respectively.

Table 3.6.1: Grade and stage demographics table

		Total dataset	low grade	high grade	stage 1	stage 2	stage 3	stage 4
total sample	N	970	355	615	361	74	405	130
	median age (IQR)	61.05 (11.65)	60.57 (12.01)	61.33 (11.43)	60.76 (12.33)	61.42 (11.77)	61.63 (11.15)	59.86 (11.13)
	gender N (f/m)	699/271	254/101	445/170	239/122	43/31	334/71	83/47
endometrial	N	247	150	97	164	20	50	13
	median age (IQR)	63.04 (11.09)	61.51 (11.04)	65.42 (10.79)	62.69 (11.17)	66.2 (10.97)	62.5 (9.70)	64.77 (15.12)
	gender N (f/m)	247/0	150/0	97/0	164/0	20/0	50/0	13/0
renal	N	416	177	239	197	40	112	67
	median age (IQR)	60.66 (12.09)	59.84 (12.67)	61.28 (11.64)	59.16 (13.03)	59.08 (11.33)	63.96 (11.46)	60.51 (9.58)
	gender N (f/m)	145/271	76/101	69/170	75/122	09/31	41/71	20/47
ovarian	N	307	28	279	0	14	243	50
	median age (IQR)	59.98 (11.30)	60.14 (12.75)	59.96 (11.17)	N/A	61.29 (12.82)	60.37 (11.14)	57.72 (11.61)
	gender N (f/m)	307/0	28/0	279/0	0/0	14/0	243/0	50/0

The number of samples, the number of each gender, and the sample age median and interquartile range (IQR) for the total dataset, and the total dataset when stratified by cancer grade (low-grade / high grade) and cancer stage (stage 1 / stage 2 / stage 3 / stage 4).

Table 3.6.2: Descriptive statistics and class imbalance tests across and within cancers

			Total sample	low/high	Wilcoxon rank sum test across low and high grade	stages I / II / III / IV	Kruskal-Wallis test across cancer stage
total sample	variant frequencies (median/inter-quartile range)	frame shift del	2 (3)	3/2 (4/4)	W= 134143.5, p=1.79×10^{-9}	3/3/1/1 (5/5/3/3)	$\chi^2(3) = 86.96, p = 9.87 \times 10^{-19}$
		Frame_Shift_Ins	1 (2)	1/1 (2/2)	W= 131070.5, p=4.3×10^{-8}	1/1/0/1 (3/2/1/1)	$\chi^2(3) = 67.56, p = 1.42 \times 10^{-14}$
		in-frame del	1 (2)	1/0 (2/1)	W= 128290, p=1.23×10^{-6}	1/1/0/0 (2/2/1/1)	$\chi^2(3) = 74.86, p = 3.89 \times 10^{-16}$
		In_Frame_Ins	0 (0)	0/0 (1/0)	W= 119427, p=1.1×10^{-3}	0/0/0/0 (1/0/0/0)	$\chi^2(3) = 18.28, p = 3.9 \times 10^{-4}$
		Missense_Mutation	38 (27)	38/38 (28/27)	W= 112397.5, p= 0.44	39/43/36/36 (32/25.25/23/28.5)	$\chi^2(3) = 15.08, p = 1.0 \times 10^{-3}$
		Nonsense_Mutation	3 (3.75)	3/3 (3/3)	W= 120459.5, p= 6.63×10^{-3}	3/3/3/3 (4/3/3/2)	$\chi^2(3) = 26.65, p = 2.0 \times 10^{-5}$
		nonstop mutation	0 (0)	0/0 (0/0)	W= 110266, p= 0.68	0/0/0/0 (0/0/0/0)	$\chi^2(3) = 20.67, p = 1.20 \times 10^{-4}$
		RNA	1 (5)	2/1 (5/4)	W= 134588, p= 0	3/2.5/0/1 (5/6.75/2/4)	$\chi^2(3) = 111.09, p = 6.40 \times 10^{-24}$
		Silent	12 (10)	13/11 (11/9)	W= 122058, p= 2.13×10^{-3}	13/12.5/11/11 (11/11.75/9/9.75)	$\chi^2(3) = 30.06, p = 1.34 \times 10^{-6}$
		Splice_Site	1 (2)	1/1 (2/2)	W= 120098, p= 7.48×10^{-3}	1/1/1/1.5 (2/2/2/2)	$\chi^2(3) = 16.75, p = 8.0 \times 10^{-4}$
	Transition/Transversion SNV frequency (median/inter-quartile range)	A>C/T>G	4 (4)	3/4 (4/4)	W= 106307, p= 0.49	4/5/3/4 (5/4/4/4)	$\chi^2(3) = 1.76, p = 5.30 \times 10^{-4}$
		A>G/T>C	8 (8)	8/8 (10/7)	W= 118105.5, p= 0.03	9/8/7/8 (11/9.75/7/7.75)	$\chi^2(3) = 3.40, p = 1.98 \times 10^{-7}$
		A>T/T>A	4 (6)	4/4 (5/6)	W= 104993, p= 0.32	5/5/4/4.5 (6/5.75/5/6)	$\chi^2(3) = 4.67, p = 0.20$
		C>A/G>T	10 (9)	10/10 (9/9)	W= 112263.5, p= 0.46	10/12/9/10 (11/9.75/9/8)	$\chi^2(3) = 1.08, p = 0.01$
		C>G/G>C	7 (6)	6/7 (4.5/8)	W= 90722, p= 1.0×10^{-5}	6/8/7/7 (6/7/7/7)	$\chi^2(3) = 1.39, p = 3.12 \times 10^{-3}$
		C>T/G>A	21 (14)	22/20 (19/13)	W= 127045.5, p= 2.0×10^{-5}	24/22.5/20/18.5 (19/13.5/12/13.75)	$\chi^2(3) = 4.06, p = 7.76 \times 10^{-9}$
			Total sample	low/high	Wilcoxon rank sum test across low and high grade	stages I / II / III / IV	Kruskal-Wallis test across cancer stage
endometrial	variant frequencies (median/inter-quartile range)	frame shift del	3 (3)	3/3 (4/5)	W= 6971.5, p= 0.58	3/3/3/2 (5/5.5/2/2)	$\chi^2(3) = 3.26, p = 0.35$
		Frame_Shift_Ins	1 (2)	1/1 (2/2)	W= 7402.5, p= 0.81	1/1/1/1 (2/2/2/2)	$\chi^2(3) = 0.62, p = 0.89$
		in-frame del	1 (2)	1/2 (2/3)	W= 6608.5, p= 0.22	2/1.5/1.5/1 (3/4/2.75/4)	$\chi^2(3) = 0.63, p = 0.89$
		In_Frame_Ins	0 (0)	0/0 (1/1)	W= 6808.5, p= 0.33	0/0/0.5/0 (1/1/1/1)	$\chi^2(3) = 1.821, p = 0.61$
		Missense_Mutation	45 (27)	41/52 (164.25/235)	W= 6470, p= 0.14	47/58.5/40/40 (218.25/144.75/162.5/150)	$\chi^2(3) = 1.68, p = 0.64$
		Nonsense_Mutation	4 (3.75)	4/5 (12/17)	W= 7016, p= 0.64	4/7.5/4/3 (15.25/13.5/10/9)	$\chi^2(3) = 4.41, p = 0.22$
		nonstop mutation	0 (0)	0/0 (0/1)	W= 6346.5, p= 0.02	0/0/0/0 (0/0/0/0)	$\chi^2(3) = 2.07, p = 0.56$
		RNA	1 (5)	1/2 (4/4)	W= 7054, p= 0.68	2/2/1/1 (4/3.5/3/2)	$\chi^2(3) = 3.63, p = 0.30$
		Silent	14 (10)	14/16 (65.25/95)	W= 6652, p= 0.26	15/19/12.5/12 (82.25/58.5/72.75/47)	$\chi^2(3) = 1.74, p = 0.63$
		Splice_Site	2 (2)	2/3 (4.75/5)	W= 6450, p= 0.12	2/3/1/3 (6/5/3.75/3)	$\chi^2(3) = 2.16, p = 0.54$
	Transition/Transversion SNV frequency (median/inter-quartile range)	A>C/T>G	2 (5)	2/3 (4/5)	W= 6171.5, p= 0.04	2/3.5/3/2 (5.25/3.25/4.75/5)	$\chi^2(3) = 1.22, p = 0.75$
		A>G/T>C	7 (41)	6/8 (35.5/49)	W= 6387, p= 0.10	6/8/7/5 (47.75/35.75/35.5/11)	$\chi^2(3) = 2.88, p = 0.41$
		A>T/T>A	3 (7)	2/4 (6/8)	W= 6054, p= 0.02	3/2/3.5/2 (7/7/4.75/3)	$\chi^2(3) = 0.63, p = 0.89$
		C>A/G>T	10 (33)	9/12 (32/42)	W= 6337, p= 0.09	10/16.5/8.5/12 (36/24.75/26.5/26)	$\chi^2(3) = 0.63, p = 0.89$
		C>G/G>C	6 (6.5)	5/8 (5/10)	W= 4918.5, p= 2.0×10^{-5}	5/7/6/6 (5/11.25/5.75/11)	$\chi^2(3) = 5.91, p = 0.12$
		C>T/G>A	36 (190.5)	38/35 (176.25/230)	W= 7270.5, p= 0.99	39/37.5/32/33 (221.75/157.5/169.25/101)	$\chi^2(3) = 3.19, p = 0.36$
			Total sample	low/high	Wilcoxon rank sum test across low and high grade	stages I / II / III / IV	Kruskal-Wallis test across cancer stage
kidney	variant frequencies (median/inter-quartile range)	frame shift del	3 (3)	3/3 (5/4)	W= 22394, p= 0.30	3/4/3/2 (5/4/4/2)	$\chi^2(3) = 8.02, p = 0.5$
		Frame_Shift_Ins	1 (2)	1/1 (2/3)	W= 21767.5, p= 0.60	1/2/1/1 (3/2/3/2)	$\chi^2(3) = 3.95, p = 0.267$
		in-frame del	1 (2)	1/1 (2/2)	W= 22110.5, p= 0.40	1/1/1/0 (2/2/2/1)	$\chi^2(3) = 8.07, p = 0.04$
		In_Frame_Ins	0 (0)	0/0 (1/1)	W= 21515, p= 0.70	0/0/0/0 (1/0/1/0)	$\chi^2(3) = 2.23, p = 0.53$
		Missense_Mutation	38 (27)	36/39 (18/22.5)	W= 18502.5, p= 0.03	36/40/39/35 (18/21.75/17/24)	$\chi^2(3) = 4.61, p = 0.20$
		Nonsense_Mutation	3 (3.75)	3/3 (3/2)	W= 19769.5, p= 0.25	3/2/3/2 (2/2.25/2/2)	$\chi^2(3) = 0.52, p = 0.92$
		nonstop mutation	0 (0)	0/0 (0/0)	W= 20616.5, p= 0.47	0/0/0/0 (0/0/0/0)	$\chi^2(3) = 0.90, p = 0.83$
		RNA	4 (5)	4/4 (5/6)	W= 22598, p= 0.23	4/6/4/3 (5/5.25/6/5)	$\chi^2(3) = 4.57, p = 0.21$
		Silent	12 (10)	12/12 (9/8)	W= 20614, p= 0.66	12/11/13/12 (8/8/7.25/8.5)	$\chi^2(3) = 2.94, p = 0.40$
		Splice_Site	1 (2)	1/1 (1/2)	W= 23263.5, p= 0.07	1/1/1/2 (1/2/2/1.5)	$\chi^2(3) = 1.357, p = 0.72$
	Transition/Transversion SNV frequency (median/interquartile range)	A>C/T>G	5 (4.25)	4/5 (4/5)	W= 18928, p= 0.07	5/5/5/5 (4/5/5/4)	$\chi^2(3) = 1.84, p = 0.61$
		A>G/T>C	10 (7)	9/10 (7/7)	W= 20345, p= 0.51	10/9.5/10/10 (7/8.5/7.25/6)	$\chi^2(3) = 0.25, p = 0.97$
		A>T/T>A	6 (4)	5/6 (3/6)	W= 19989.5, p= 0.34	6/6/6/6 (4/5.25/4.25/6.5)	$\chi^2(3) = 1.40, p = 0.71$
		C>A/G>T	11 (8)	10/11 (7/7)	W= 19242, p= 0.11	9/11/13/11 (7/7/6/7)	$\chi^2(3) = 10.37, p = 0.02$
		C>G/G>C	7 (5)	6/7 (4/6)	W= 20045.5, p= 0.36	7/7/7/5 (5/6/5/6)	$\chi^2(3) = 2.20, p = 0.53$
		C>T/G>A	19 (11)	18/20 (9/12)	W= 19302.5, p= 0.13	19/19.5/20/18 (10/8.5/10.25/12.5)	$\chi^2(3) = 0.91, p = 0.82$
			Total sample	low/high	Wilcoxon rank sum test across low and high grade	stages I / II / III / IV	Kruskal-Wallis test across cancer stage
ovarian	variant frequencies (median/inter-quartile range)	frame shift del	1 (3)	1/1 (2/2)	W= 4293, p= 0.37	NA/1/1/1 (NA/2.5/2/1)	$\chi^2(3) = 3.1, p = 0.2$
		Frame_Shift_Ins	0 (2)	0/0 (1/1)	W= 4091.5, p= 0.61	NA/0/0/0 (NA/1/1/0)	$\chi^2(3) = 2.32, p = 0.31$
		in-frame del	0 (2)	0/0 (1/1)	W= 3786, p= 0.75	NA/1/0/0 (NA/1/1/1)	$\chi^2(3) = 4.37, p = 0.11$
		In_Frame_Ins	0 (0)	0/0 (0/0)	W= 3831.5, p= 0.74	NA/0/0/0 (NA/0/0/0)	$\chi^2(3) = 0.644, p = 0.72$
		Missense_Mutation	33 (27)	34.5/33 (12.75/25.5)	W= 4147.5, p= 0.59	NA/41.5/32/36 (NA/19/24/28.5)	$\chi^2(3) = 7.42, p = 0.02$
		Nonsense_Mutation	2 (3.75)	2.5/2 (3/3)	W= 4042.5, p= 0.76	NA/3/2/2.5 (NA/3.5/3/3)	$\chi^2(3) = 0.98, p = 0.61$
		nonstop mutation	0 (0)	0/0 (0/0)	W= 3710, p= 0.23	NA/0/0/0 (NA/0/0/0)	$\chi^2(3) = 3.85, p = 0.15$
		RNA	0 (5)	0/0 (1/1)	W= 3828, p= 0.83	NA/0/0/0 (NA/0/1/1)	$\chi^2(3) = 3.54, p = 0.17$
		Silent	10 (10)	12/9 (6.75/8)	W= 4314.5, p= 0.36	NA/14.5/9/10 (NA/10.5/8/10)	$\chi^2(3) = 8.54, p = 0.01$
		Splice_Site	1 (2)	1/1 (2/2)	W= 3808.5, p= 0.82	NA/2/1/1 (NA/1/2/2)	$\chi^2(3) = 2.55, p = 0.28$
	Transition/Transversion SNV frequency (median/interquartile range)	A>C/T>G	3 (4)	3/3 (2.25/4)	W= 3959, p= 0.91	NA/5.5/3/3 (NA/3.75/4/4)	$\chi^2(3) = 8.71, p = 0.01$
		A>G/T>C	6 (5)	6.5/6 (6/5)	W= 4088, p= 0.68	NA/7/6/7 (NA/9/4.5/6)	$\chi^2(3) = 4.88, p = 0.09$
		A>T/T>A	4 (4)	3.5/4 (5.5/4)	W= 3863.5, p= 0.92	NA/5/3/3.5 (NA/3.75/4/5.75)	$\chi^2(3) = 4.03, p = 0.13$
		C>A/G>T	8 (8)	8.5/8 (8.25/8)	W= 4276, p= 0.41	NA/12/8/8.5 (NA/8/8/6.75)	$\chi^2(3) = 7.25, p = 0.03$
		C>G/G>C	8 (9)	8/7 (7/9.5)	W= 4020.5, p= 0.80	NA/12/7/9 (NA/3.5/8.5/9.5)	$\chi^2(3) = 8.58, p = 0.01$
		C>T/G>A	18 (11)	19/18 (12.25/11)	W= 4084, p= 0.69	NA/20/18/18 (NA/13/11/13.75)	$\chi^2(3) = 1.19, p = 0.55$

The descriptive statistics, and inferential statistics for each variant classification, and transition and transversion variant type across all three cancers and within each cancer.

Table 3.6.3: Stage assignment frequencies stratified by grade

	Total			renal cell carcinoma			ovarian carcinoma			endometrial carcinoma		
	Total	High grade	Low grade	Total	High grade	Low grade	Total	High grade	Low grade	Total	High grade	Low grade
IIA	2	1	1	0	0	0	2	1	1	0	0	0
IIB	4	4	0	0	0	0	4	0	4	0	0	0
IIC	8	3	5	0	0	0	8	5	3	0	0	0
IIIA	3	2	1	0	0	0	3	1	2	0	0	0
IIIB	14	13	1	0	0	0	14	1	13	0	0	0
IIIC	226	213	13	0	0	0	226	13	213	0	0	0
IV	50	43	7	0	0	0	50	7	43	0	0	0
StageI	198	74	124	197	123	74	0	0	0	1	1	0
StageIA	83	20	63	0	0	0	0	0	0	83	63	20
StageIB	70	20	50	0	0	0	0	0	0	70	50	20
StageIC	10	4	6	0	0	0	0	0	0	10	6	4
StageII	51	30	21	40	16	24	0	0	0	11	5	6
StageIIA	4	3	1	0	0	0	0	0	0	4	1	3
StageIIB	5	1	4	0	0	0	0	0	0	5	4	1
StageIII	112	83	29	112	29	83	0	0	0	0	0	0
StageIIIA	17	4	13	0	0	0	0	0	0	17	13	4
StageIIIC	14	10	4	0	0	0	0	0	0	14	4	10
StageIIIC1	8	7	1	0	0	0	0	0	0	8	1	7
StageIIIC2	11	10	1	0	0	0	0	0	0	11	1	10
StageIV	69	59	10	67	9	58	0	0	0	2	1	1
StageIVA	1	1	0	0	0	0	0	0	0	1	0	1
StageIVB	10	10	0	0	0	0	0	0	0	10	0	10

The rows denote the stage classifications in the Pan-Cancer clinical data across endometrial carcinoma, ovarian carcinoma and renal cell carcinoma. The columns indicate the frequency of each of the cancer stages in total, and stratified by broad cancer grade, and cancer type.

Table 3.6.4: Grade assignment stratified by tumour stage (low/high)

	Total			renal cell carcinoma			ovarian carcinoma			endometrial carcinoma		
	Total	High-stage	Low-stage	Total	High-stage	Low-stage	Total	High-stage	Low-stage	Total	High-stage	Low-stage
G1	7	0	7	7	7	0	0	0	0	0	0	0
G2	198	60	138	170	132	38	28	6	22	0	0	0
G3	450	354	96	172	88	84	278	8	270	0	0	0
G4	68	58	10	67	10	57	1	0	1	0	0	0
Grade1	76	10	66	0	0	0	0	0	0	76	66	10
Grade2	74	10	64	0	0	0	0	0	0	74	64	10
Grade3	92	40	52	0	0	0	0	0	0	92	52	40
HighGrade	5	3	2	0	0	0	0	0	0	5	2	3

The rows denote the cancer grade classifications in the Pan-Cancer clinical data across endometrial carcinoma, ovarian carcinoma and renal cell carcinoma. The columns indicate the frequency of each of the cancer grades in total, and stratified by broad cancer stage, and cancer type.

Table 3.6.5: Complete across cancer grade classification models

		Odds Ratio (OR)	CI (2.5%, 97.5%)	p-value
model AG	(Intercept)	1.00	(0.84,1.20)	1
model AGP	(Intercept)	0.43	(0.27, 0.65)	8.68×10^{-5}
model AGPV	gender (male)	2.22	(1.29, 3.87)	4.43×10^{-3}
	TP53	12.25	(6.90, 22.47)	2.0×10^{-16}
	ARID1A	2.13	(0.97, 4.70)	0.06
	CTCF	0.36	(0.08, 1.18)	0.13
	CTNNB1	0.09	(0.01, 0.33)	1.69×10^{-3}
	*PIK3*R1	0.41	(0.15, 1.04)	0.069
model AGV	(Intercept)	0.98	(0.70,1.36)	0.88
	'C>G/G>C'	1.06	(1.03,1.10)	8.21×10^{-4}
	frame shift del	0.88	(0.83,0.93)	6.13×10^{-6}
model AGS	(Intercept)	0.30	(0.21,0.42)	2.79×10^{-12}
model AGSV	stage (high)	9.38	(6.18,14.48)	2.0×16
	gender (male)	1.45	(0.90,2.34)	0.13
model AGSP	(Intercept)	0.24	(0.15,0.35)	1.34×10^{-11}
model AGSPV	stage (high)	4.60	(2.88,7.43)	2.64×10^{-10}
	gender (male)	2.23	(1.33,3.78)	2.52×10^{-3}
	TP53	6.32	(3.56,11.44)	5.35×10^{-10}
	CTNNB1	0.10	(0.02,0.38)	3.45×10^{-3}
model AGT	(Intercept)	0.15	(0.04,0.52)	2.88×10^{-3}
model AGTV	age	1.02	(0.1,1.03)	0.11
	gender (male)	2.73	(1.47,5.19)	1.72×10^{-3}
	tumour type (OV)	14.86	(7.53,30.66)	4.19×10^{-14}
	tumour type (UCEC)	0.98	(0.52,1.88)	0.95
model AGTP	(Intercept)	0.37	(0.22,0.62)	1.95×10^{-4}
model AGTPV	gender (male)	2.56	(1.38,4.86)	3.29×10^{-3}
	tumour type (OV)	2.46	(0.95,6.36)	0.06
	tumour type (UCEC)	0.62	(0.31,1.25)	0.18
	TP53	7.40	(3.60,15.73)	8.84×10^{-08}
model AGST	(Intercept)	0.12	(0.03,0.42)	1.25×10^{-3}
model AGSTV	age	1.01	(0.99,1.03)	0.33
	stage (high)	4.55	(2.78,7.52)	2.27×10^{-09}
	gender (male)	2.75	(1.43,5.47)	3.03×10^{-3}
	tumour type (OV)	6.36	(2.99,13.92)	2.22×10^{-6}
	tumour type (UCEC)	1.24	(0.63,2.48)	0.54
model AGSTP	(Intercept)	0.18	(0.11,0.26)	2.0×10^{-16}
model AGSTPV	stage (high)	4.89	(3.08,7.84)	2.61×10^{-11}
	gender (male)	2.89	(1.74,4.84)	4.76×10^{-05}
	TP53	7.61	(4.35,13.64)	2.91×10^{-12}

A= age, **G**= gender, **S**= Stage, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency

Table 3.6.6: Complete across cancer stage classification models

		Odds Ratio (OR)	CI (2.5%, 97.5%)	p-value
model AG	(Intercept)	1.2	(0.99,1.46)	0.07
	gender (male)	0.54	(0.37,0.77)	7.80×10^{-4}
model AGP	(Intercept)	0.68	(0.54,0.87)	1.74×10^{-3}
	TP53	10.05	(6.45,16.13)	2.0×10^{-16}
	ARID1A	0.52	(0.20,1.23)	0.15
	CHD3	0.27	(0.04,1.19)	0.12
	PIK3CA	0.26	(0.12,0.51)	1.5×10^{-4}
	PIK3R1	0.24	(0.09,0.55)	1.58×10^{-3}
model AGV	(Intercept)	2.12	(1.63,2.77)	2.56×10^{-8}
	gender (male)	0.50	(0.34,0.73)	4.37×10^{-4}
	frame shift del	0.91	(0.85,0.97)	4.32×10^{-3}
	in-frame del	0.85	(0.75,0.94)	4.41×10^{-3}
	nonstop mutation	0.86	(0.68,1.01)	0.10
model AGPV	(Intercept)	0.84	(0.61,1.14)	0.25
	frame shift del	0.95	(0.89,1.00)	0.06
	TP53	8.94	(5.71,14.36)	2.0×10^{-16}
	ARID1A	0.52	(0.20,1.23)	0.15
	PIK3CA	0.24	(0.12,0.47)	4.75×10^{-5}
	PIK3R1	0.24	(0.09,0.56)	1.72×10^{-3}
model AGGr	(Intercept)	0.33	(0.23,0.45)	2.67×10^{-11}
	grade (high)	9.16	(6.16,13.85)	2.0×10^{-16}
	gender (male)	0.40	(0.26,0.60)	1.35×10^{-5}
model AGGrP	(Intercept)	0.37	(0.24,0.57)	5.15×10^{-6}
	grade (high)	4.72	(3.00,7.52)	3.33×10^{-11}
	gender (male)	0.59	(0.35,0.98)	4.39×10^{-2}
	TP53	5.40	(3.12,9.48)	2.67×10^{-9}
	FNDC1	0.16	(0.03,0.69)	1.97×10^{-2}
	PIK3CA	0.21	(0.10,0.42)	1.88×10^{-5}
	PIK3R1	0.24	(0.09,0.57)	2.06×10^{-3}
model AGGrV	(Intercept)	0.56	(0.38,0.82)	3.08×10^{-3}
model AGGrPV	grade (high)	8.30	(5.53,12.68)	2.0×10^{-16}
	gender (male)	0.39	(0.25,0.60)	1.59×10^{-5}
	frame shift del	0.92	(0.854,0.982)	0.01
	in-frame del	0.90	(0.79,1.01)	0.09
	nonstop mutation	0.85	(0.68,1.018)	0.11
model AGT	(Intercept)	0.20	(0.07,0.59)	3.44×10^{-3}
model AGTP	age	1.02	(1.00,1.04)	0.04
model AGTV	tumour type (OV)	27.31	(14.37,57.72)	2.0×10^{-16}
model AGTPV	tumour type (UCEC)	0.41	(0.25,0.65)	2.02×10^{-4}
model AGGrT	(Intercept)	0.10	(0.03,0.32)	9.8×10^{-5}
model AGGrTP	age	1.01	(0.10,1.03)	0.12
model AGGrTV	grade (high)	4.50	(2.89,7.11)	5.14×10^{-11}
model AGGrTPV	tumour type (OV)	20.50	(10.6,43.87)	2.0×10^{-16}
	tumour type (UCEC)	0.51	(0.30,0.84)	8.91×10^{-3}

A= age, **G**= gender, **Gr**= grade, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency

Table 3.6.7: Cancer grade prediction models within cancers

			Odds Ratio (OR)	CI (2.5%, 97.5%)	p-value
endometrial carcinoma	model WA	(Intercept)	0.07	(0.01,0.58)	0.017
	model WAV	age	1.04	(1.01,1.08)	0.016
	model WAP	(Intercept)	0.92	(0.43,1.97)	0.84
	model WAPV	*PTEN*	0.37	(0.15,0.90)	0.03
		TP53	13.05	(4.39,48.71)	1.87×10^{-5}
	model WAGS	(Intercept)	0.05	(0.00,0.42)	8.05×10^{-3}
	model WAGSV	age	1.05	(1.01,1.08)	0.02
		stage (high)	3.37	(1.49,8.04)	4.39×10^{-3}
	model WAGSP	(Intercept)	0.77	(0.34,1.71)	0.52
	model WAGSPV	stage (high)	2.06	(0.77,5.57)	0.15
		PTEN	0.38	(0.15,0.94)	0.04
		TP53	11.78	(3.91,44.28)	4.64×10^{-5}
			Odds Ratio (OR)	CI (2.5%, 97.5%)	p-value
renal cell	model WA	(Intercept)	1.00	(0.78,1.29)	1
	model WAP				
	model WAV				
	model WAPV				
	model WAG	(Intercept)	1.00	(0.78,1.29)	1
	model WAGP				
	model WAGV				
	model WAGPV				
	model WAGS	(Intercept)	0.40	(0.23,0.66)	5.42×10^{-4}
	model WAGSP	stage (high)	5.05	(2.89,9.06)	2.69×10^{-8}
	model WAGSV	gender (male)	1.53	(0.87,2.73)	0.15
	model WAGSPV				
			Odds Ratio (OR)	CI (2.5%, 97.5%)	p-value
ovarian carcinoma	model WA	(Intercept)	1.00	(0.53,1.)	1
	model WAP				
	model WAV				
	model WAPV				
	model WAGS	(Intercept)	0.00	(NA,7.76×10^{90})	0.99
	model WAGSP	stage (high)	53890100.00	(0,NA)	0.99
	model WAGSV				
	model WAGSPV				

W= within cancer, A= age, **G**= gender, **S**= Stage, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency

Table 3.6.8: Cancer stage prediction models within cancers

			Odds Ratio (OR)	CI (2.5%, 97.5%)	p-value
	model WA	**(Intercept)**	1.00	0	1
	model WAP				
	model WAV				
	model WAPV				
endometrial					
carcinoma	**model WAGGr**	**(Intercept)**	0.37	-2.68	7.29×10^{-3}
	model WAGGrP	**grade (high)**	5.76	3.61	3.04×10^{-4}
	model WAGGrV				
	model WAGGrPV				

			Odds Ratio (OR)	CI (2.5%, 97.5%)	p-value
	model WA	**(Intercept)**	0.35	(0.09,1.25)	0.11
	model WAP	**age**	1.02	(1.0,1.04)	0.10
	model WAV				
	model WAPV				
	model WAG	**(Intercept)**	0.35	(0.09,1.25)	0.11
renal	**model WAGP**	**age**	1.02	(1.00,1.04)	0.10
cell	**model WAGV**				
	model WAGPV				
	model WAGGr	**(Intercept)**	0.12	(0.03,0.49)	4.23×10^{-3}
	model WAGGrP	**age**	1.02	(0.99,1.04)	0.17
	model WAGGrV	**grade (high)**	5.98	(3.34,11.06)	4.41×10^{-9}
	model WAGGrPV				

			Odds Ratio (OR)	CI (2.5%, 97.5%)	p-value
	model WA	**(Intercept)**	1.00	(0.39,2.56)	1
	model WAP				
	model WAV				
	model WAPV				
ovarian					
carcinoma	**model WAGGr**	**(Intercept)**	0.00	(NA,2.33×10^{154})	1.00
	model WAGGrP	**grade (high)**	208167800.000	(0,NA)	1.00
	model WAGGrV				
	model WAGGrPV				

W= within cancers, **A**= age, **G**= gender, **Gr**= grade, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency

Table 3.6.9: Endometrial carcinoma grade and stage classification statistics

		Sensitivity	Specificity	Accuracy	Positive predictive value (PPV)	Area Under the Curve (AUC)
Grade	**model WA**	0.59	0.62	0.62	0.37	0.61
	model WAP	0.69	0.92	0.86	0.76	0.83
	model WAV	0.50	0.50	0.50	NA	0.50
	model WAPV	0.69	0.92	0.86	0.76	0.83
	model WAGS	0.66	0.68	0.68	0.44	0.75
	model WAGSP	0.75	0.91	0.86	0.75	0.88
	model WAGSV	0.66	0.68	0.68	0.44	0.75
	model WAGSPV	0.75	0.91	0.86	0.27	0.88
Stage	**model WA**	0.50	0.50	0.50	NA	0.50
	model WAP	0.50	0.50	0.50	NA	0.50
	model WAV	0.50	0.50	0.50	NA	0.50
	model WAPV	0.50	0.50	0.50	NA	0.50
	model WAGGr	0.52	0.73	0.70	0.22	0.63
	model WAGGrP	0.52	0.73	0.70	0.22	0.63
	model WAGGrV	0.52	0.73	0.70	0.22	0.63
	model WAGGrPV	0.52	0.73	0.70	0.22	0.63

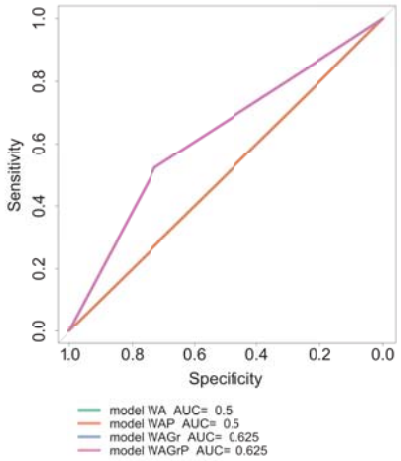
W= within cancers, **A**= age, **G**= gender, **Gr**= grade, **S**= Stage, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency

Table 3.6.10: Renal cell carcinoma grade and stage classification statistics

		Sensitivity	Specificity	Accuracy	Positive predictive value (PPV)	Area Under the Curve (AUC)
Grade	**model WA**	0.50	0.50	0.50	NA	0.50
	model WAP	0.50	0.50	0.50	NA	0.50
	model WAV	0.50	0.50	0.50	NA	0.50
	model WAPV	0.50	0.50	0.50	NA	0.50
	model WAG	0.50	0.50	0.50	NA	0.50
	model WAGP	0.50	0.50	0.50	NA	0.50
	model WAGV	0.50	0.50	0.50	NA	0.50
	model WAGPV	0.50	0.50	0.50	NA	0.50
	model WAGS	0.59	0.81	0.66	0.87	0.75
	model WAGSP	0.59	0.81	0.66	0.87	0.75
	model WAGSV	0.59	0.81	0.66	0.87	0.75
	model WAGSPV	0.59	0.81	0.66	0.87	0.75
Stage	**model WA**	0.55	0.63	0.60	0.43	0.60
	model WAP	0.50	0.50	0.50	NA	0.50
	model WAV	0.50	0.50	0.50	NA	0.50
	model WAPV	0.50	0.50	0.50	NA	0.50
	model WAG	0.55	0.63	0.60	0.43	0.60
	model WAGP	0.55	0.63	0.60	0.43	0.60
	model WAGV	0.55	0.63	0.60	0.43	0.60
	model WAGPV	0.55	0.63	0.60	0.43	0.60
	model WAGGr	0.72	0.61	0.65	0.48	0.72
	model WAGGrP	0.72	0.61	0.65	0.48	0.72
	model WAGGrV	0.72	0.61	0.65	0.48	0.72
	model WAGGrPV	0.72	0.61	0.65	0.48	0.72

W= within cancers, **A**= age, **G**= gender, **Gr**= grade, **S**= Stage, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency

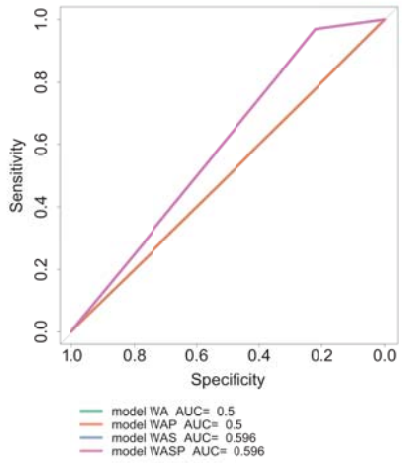
Figure 3.6.2: Endometrial carcinoma stage classification ROC



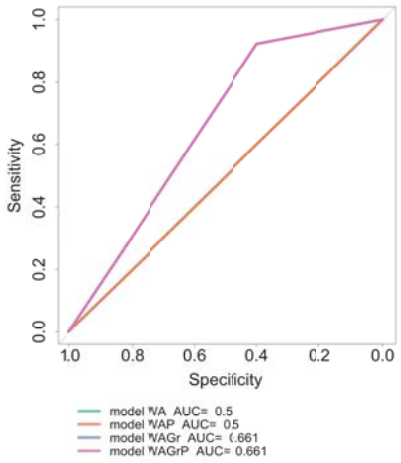
W= within cancers, **A**= age, **Gr**= grade, **P**= proteins. Both the WA, and WAP models were uninformative. By including grading information, stage prediction was improved in model WAGr.

Figure 3.6.3: Ovarian carcinoma ROC curves

(a) ovarian carcinoma grade classification baseline plus protein features



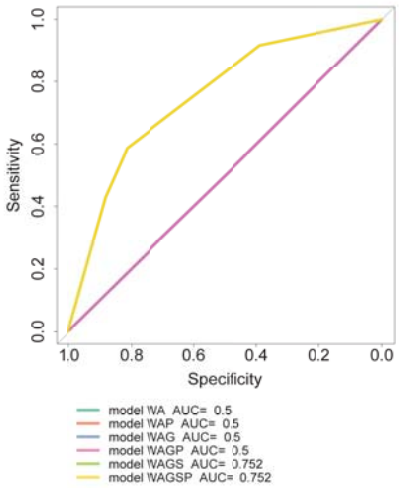
(b) ovarian carcinoma stage classification plus protein features



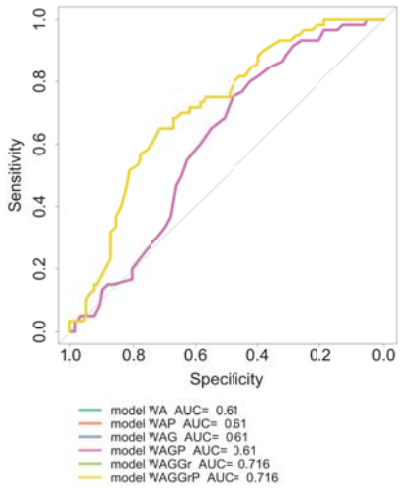
W= within cancers, **A**= age, **Gr**= grade, **S**= Stage, **P**= proteins. Models WA, and WAP were non-informative when predicting grade, and stage. By adding stage and grade features the prediction of cancer stage and grade respectively was improved. Where protein features were added to the analysis (models WAP, WASP, and WAGrP) no protein features were included in final models after AIC-based model selection.

Figure 3.6.4: Renal carcinoma stage classification ROC curves

(a) renal carcinoma grade classification baseline plus protein features



(b) renal carcinoma stage classification baseline plus protein features



W= within cancers, **A**= age, **G**= gender, **Gr**= grade, **S**= Stage, **P**= proteins. Models WA, WAG, WAP, and WAGP were non-informative when predicting grade, and stage. By adding stage and grade features the prediction of cancer stage and grade respectively was improved. Where protein features were added to the analysis (models WAP, WAGP, WAGSP, WASP, and WAGGrP) no protein features were included in final models after AIC-based model selection.

Table 3.6.11: Ovarian carcinoma grade and stage classification statistics

		Sensitivity	Specificity	Accuracy	Positive predictive value (PPV)	Area Under the Curve (AUC)
Grade	**model WA**	0.50	0.50	0.50	NA	0.50
	model WAP	0.50	0.50	0.50	NA	0.50
	model WAV	0.50	0.50	0.50	NA	0.50
	model WAPV	0.50	0.50	0.50	NA	0.50
	model WAGS	0.97	0.22	0.94	0.97	0.60
	model WAGSP	0.97	0.22	0.94	0.97	0.60
	model WAGSV	0.97	0.22	0.94	0.97	0.60
	model WAGSPV	0.97	0.22	0.94	0.97	0.60
Stage	**model WA**	0.50	0.50	0.50	NA	0.50
	model WAP	0.50	0.50	0.50	NA	0.50
	model WAV	0.50	0.50	0.50	NA	0.50
	model WAPV	0.50	0.50	0.50	NA	0.50
	model WAGGr	0.92	0.40	0.91	0.99	0.66
	model WAGGrP	0.92	0.40	0.91	0.99	0.66
	model WAGGrV	0.92	0.40	0.91	0.99	0.66
	model WAGGrPV	0.92	0.40	0.91	0.99	0.66

W= within cancers, **A**= age, **G**= gender, **Gr**= grade, **S**= Stage, **T**= tumour type (endometrial, ovarian, renal), **P**= proteins, **V**= variant frequency

Chapter 4

Predicting cancer types using TCGA exome sequence data and Random Forest analysis.

The PanCancer analyses have focused mainly on finding novel subgroups of cancers (Hofree et al., 2013; Ciriello et al., 2013). I decided to discriminate between known 'high-order' cancer classes corresponding to five tissue types in the PanCancer exome sequence dataset (adenocarcinoma, squamous cell carcinoma, urothelial carcinoma, a blood cancer, and brain cancer).

In chapter 4.1 I use a Random Forest (Breiman, 2001) to identify the mutated genes that discriminate between the five cancer types based exome sequence data. Based upon this work I adopted a similar Random Forest approach in chapter 4.2 to build a classifier to predict the tissue of origin of samples in the PanCancer dataset. This approach could be used to create a tool to assign a cancer of unknown primary origin (CUP) a putative tissue of origin. In the future it may aid clinicians decide between treatment options for patients with CUP. Stage IV CUPs represent 3-5 percent of cancer diagnoses, and patient outcomes are typically poor (Pavlidis & Pentheroudakis, 2012). Although gene-expression models have been developed to predict the origin of CUPs (Tothill et al., 2005), there is no model based on whole exome sequence data.

Chapter 4.1

Using exome sequence data and Random Forest analysis to identify the functional mutation patterns of five high-order cancer types

4.1.1 Introduction

The Pan-Cancer Analysis of the Cancer Genome Atlas (TCGA) data was an important milestone for the analysis of genomic cancer data. Across twelve cancer types six different data types were collected, including tumour/normal exome sequence, gene expression microarray data, DNA methylation array data, microRNA data, copy number variant data, protein expression array data, and more than one hundred clinical variables. Standardising analysis pipelines across the 12 cancer types has allowed researchers to discover new cancer subtypes defined across cancers. Most of the cross-cancer analyses have focused on the definition of new cancer subtypes by grouping samples according to common molecular alterations including somatic single nucleotide variants (SNVs), epigenetic modifications, copy number variants and gene expression changes (Ciriello et al., 2013; Lawrence et al., 2014; Hofree et al., 2013).

Multiple approaches have been used to analyse this rich data set, often employing the integration of different data types. The Dendrix algorithm (Leiserson et al., 2013a), was applied to the exome sequence data to identify the genes that were mutually exclusively mutated across the 12 cancer types. The ENCAPP framework (Das et al., 2015) used an elastic net approach to integrate the molecular data for breast, ovarian and colorectal cancers and uncover prognostic biomarkers. Ciriello et al. (2013) used hierarchical clustering of *selected functional events* from somatic mutations, copy-number variants, and DNA methylation events to discover that the twelve Pan-Cancer cancer types could be assigned to two groups, characterised by either high numbers of SNVs, or high numbers of copy number alterations. Network-based stratification (Hofree et al., 2013) of the Pan-Cancer tumour types identified new subclasses of ovarian, uterine, and lung adenocarcinoma cancers that showed different survival patterns based upon somatic mutation profiles that had undergone a *network smoothing* step.

Relatively little attention has been paid to the differences between high-order cancer types, which may share mutations that discriminate them from other cancer types. The multitude of human cancers can be divided in to classes based on the tissue type in which

they originate. In the Pan-Cancer dataset the twelve cancers can be divided in to five classes; adenocarcinomas, squamous cell carcinomas, urothelial carcinomas, haematological cancers and cancers of the central nervous system.

Adenocarcinomas originate from secretory cells in epithelial glandular tissue. Due to the ubiquitous nature of secretory cells in humans these tumours are very diverse.

Squamous cell carcinomas also arise from epithelial cells, but unlike adenocarcinomas these tumours do not originate from secretory cells. They are also diverse in terms of histology, prognosis, aggression, and response to treatment.

Urothelial carcinomas arise in the complex transitional epithelial membranes encapsulating the lumen of organs of the urinary system.

Haematological cancers are found in the blood bone marrow and include Acute Myeloid Leukemia (AML).

Cancers of the central nervous system (CNS) include glioblastoma multiforme (GBM), which is typically highly aggressive. It is difficult to conduct histopathological analysis on this class of tumours due to their location.

The cancers that arise from similar tissue types may share somatic mutations that contribute to cancer progression and discriminate them from cancers originating from other tissue types. In this study I have taken a high-order view of the 12 cancers of the Pan-Cancer analysis. I used somatic mutation data derived from paired tumour / normal whole exome sequencing to identify the proteins carrying functional mutations that can discriminate between the five tissue types (adenocarcinomas, squamous cell carcinomas, urothelial malignancies, blood cancers, and cancers of the central nervous system).

I conducted 10 pairwise comparisons to discriminate between the five cancer types. For each comparison I split the sample with two thirds becoming a training set and one third becoming a test set. I created Random Forest (Breiman, 2001) classification models using protein features and the clinical features; age, gender and cancer stage (where available). I

tested the ability of the models derived from each comparison to discriminate between the two cancer types in the test set. In addition, I identified the protein sets that discriminated between the each of the cancer types.

4.1.2 Methods

4.1.2.1 The data set

The Mutation Annotation Files (MAFs), containing the somatic mutation information for all 12 Pan-Cancer cancers were downloaded from Sage Synapse (accession syn1710680) on October third 2013 along with the clinical data files. I retained 3079 samples for which exome-sequence, age, gender, and (where appropriate) tumour stage data were available. After I removed the gender specific adenocarcinomas (ovarian carcinoma, endometrial carcinoma, and breast invasive ductal carcinoma) from further analysis, 1739 samples remained.

Twenty three cancer stage categories were used to describe the cancer stage of the samples. I collapsed the 23 categories in to 4 stages (I, II, III and IV). Cancer stage was not available for glioblastoma, or AML, and was not included in any comparison for either of those cancers.

4.1.2.2 Definition of the five cancer high-order cancer types

The 12 cancers were assigned to five classes according to their tissue type of origin, which represented functionally similar cancers. In this study adenocarcinomas consisted of breast adenocarcinoma (BRCA), kidney renal clear-cell carcinoma (KIRC), colon and rectum adenocarcinoma (COADREAD), lung adenocarcinoma (LUAD), and ovarian carcinoma (OV). The squamous cell carcinoma group consisted of head and neck squamous cell carcinoma (HNSC), and lung squamous cell carcinoma (LUSC). Bladder urothelial carcinoma (BLCA) represented the urothelial carcinomas, Acute myeloid leukaemia (AML) represented the haematological malignancies, and glioblastoma (GBM) represented the tumours of the CNS. The numbers of samples assigned to each of the five cancer classes are shown in Table 4.1.2.1. The urothelial cancer type contained 89 samples and was considerably smaller than the other classes.

The assignment of the twelve Pan-Cancer tumour types to the five high-order classes, and the availability of cancer stage information is shown in supplementary table 4.1.6.1.

Table 4.1.2.1: High-order cancer type assignment

tissue type	cancer type	N
adenocarcinoma	colorectal adenocarcinoma	219
	renal clear cell carcinoma	417
	lung adenocarcinoma	155
squamous cell carcinoma	head and neck squamous cell carcinoma	260
	lung squamous cell carcinoma	172
urothelial carcinoma	bladder urothelial carcinoma	89
haematological malignancy	acute myeloid leukemia	152
central nervous system	glioblastoma	275

The cancer types are assigned to the tissue type with which they share a row. The frequency of samples for each cancer type is shown in column **N**.

4.1.2.3 Creating the binary mutation matrix

I removed 'silent' mutations from the MAF file. I used the union of the lists of functionally mutated proteins across all samples to generate a binary matrix M of proteins P by samples S. Each element in M, $M[p, s]$, was set to 0 if there was no protein coding mutation for protein p and sample s, or set to 1 if at least one protein coding mutation was present for protein p and sample s.

4.1.2.3.1 Mutated protein feature selection

I used a simple heuristic to reduce the number of proteins used as features in the Random Forest models. I retained a protein feature if it carried at least one protein coding mutation in five percent or more of any cancer class in the training set. The numbers of proteins taken forward for model building across all Random Forest models are shown in Table 4.1.2.2.

4.1.2.4 Training set and test set definition.

When creating classification models it is important that the model can be tested in an independent set of samples that were not used to create the model. This allows researchers to gain insight in to how the models perform on new data. For each of the ten pairwise comparisons two thirds of the samples were assigned to the training set, which I used for model building, and one third were assigned to the test set. I down-sampled the majority

cancer class to be of equal size as the minority cancer class. The majority cancer class samples removed during down-sampling were re-assigned to the test set. Table 4.1.2.2 shows the number of samples in the training and test sets across the ten Random Forest models.

Table 4.1.2.2: Number of proteins used for model building, and training and test set sizes.

Comparison	N proteins after filter	Training set size	Test set size
adenocarcinoma, squamous	1351	576	647
adenocarcinoma, urothelial	741	118	762
squamous, urothelial	910	118	403
adenocarcinoma, GBM	1236	366	700
squamous, GBM	621	366	341
urothelial, GBM	476	118	246
adenocarcinoma, leukemia	646	204	739
squamous, leukemia	824	204	380
urothelial, leukemia	462	118	123
GBM, leukemia	32	204	223

N proteins after filter: The number of proteins which were functionally mutated in five percent of the training set samples of either class and taken forward for model building.
Training set size: The number of samples used for model building. Two thirds of the minority class samples were assigned to the training set with an equal number of majority class samples.
Test set size: The number of samples assigned to the test set. No samples were used for model building and model evaluation.

4.1.2.5 Random Forest prediction models

4.1.2.5.1 Random Forest model building

The Random Forest algorithm builds an ensemble classifier using a collection, or *forest*, of decision trees. Each decision tree provides a class prediction, or vote, for each sample. For each sample, the class prediction of the Random Forest model is the modal class across all votes in the forest.

For each comparison the training samples were divided in to a *learning set* and an *out of bag* set. The learning set and out of bag *set* can be considered as an internal *training set* and *test set*. To build each tree, a bootstrapped (Efron & Gong, 1983) sample of the learning set is used. A tree is learned by splitting the set of samples based on the feature, from a set of randomly selected features, that best partitions the samples in to their respective classes. This process continues until each subset of samples after a decision are of the same class and is repeated T times to create a forest of T trees. The out of bag set is used to

measure the classification accuracy of the forest, and the importance of each feature. The out of bag samples are run through each tree in the forest. The feature importance is the mean classification accuracy of all decisions where the feature was used, and provides a summary of the ability of each feature to correctly classify the out of bag samples across the forest. Many different measures of feature importance are available including the Gini index (Breiman, 2001) and F-measure. However, I used classification accuracy to indicate feature importance because the the Gini index overestimates the importance of correlated, but uninformative features (Nicodemus & Malley, 2009).

4.1.2.5.2 Cross Validation

The success of a Random Forest model is partly dependent on the definition of the training set and test set. If sufficient noise exists in the training data, the Random Forest algorithm will model that noise. While the model will perform well when the training set is re-classified, the classification performance in new data will not reach that of the training set. The Random Forest model will be *over-fitted* to the training set. I have tried to measure, and alleviate any over-fitting of the Random Forest models by using k-fold cross-validation.

The training set was divided in to k equally sized sub-samples called folds. A single fold is used as an out-of-bag set in order to measure the performance of the model constructed using the remaining folds. This process is performed k times in total, where each fold is treated as the out-of-bag set once, and all remaining folds are treated as the learning set; In this study $k = 10$. After using a k-fold cross-validation method, random noise resulting from the definition of each fold may still cause some over-fitting of the final model. Any remaining bias in the final classification models may be alleviated by repeating the ten-fold cross-validation using alternative folds. I repeated the ten-fold cross validation five times. By randomly assigning samples to ten folds for each cross validation iteration, the result was a 5 x ten-fold cross validation process. The performance of the Random Forest model building process was measured as the average classification accuracy across the resultant 50 classification models.

4.1.2.5.3 Recursive Feature Elimination

By using a *recursive feature elimination* step I was able to identify the number of features that lead to the best performing Random Forest model, and which were the best performing features. The recursive feature elimination step worked by assigning a number of features $f \in \{1, 2, 3, \cdots, 100, 200\}$ to be used for model building in each cross-validation run. Within each cross-validation run, the top f performing features were used to create a Random Forest model, the classification accuracy of which was used to select best performing f across all 50 cross-validation runs as measured by mean classification accuracy.

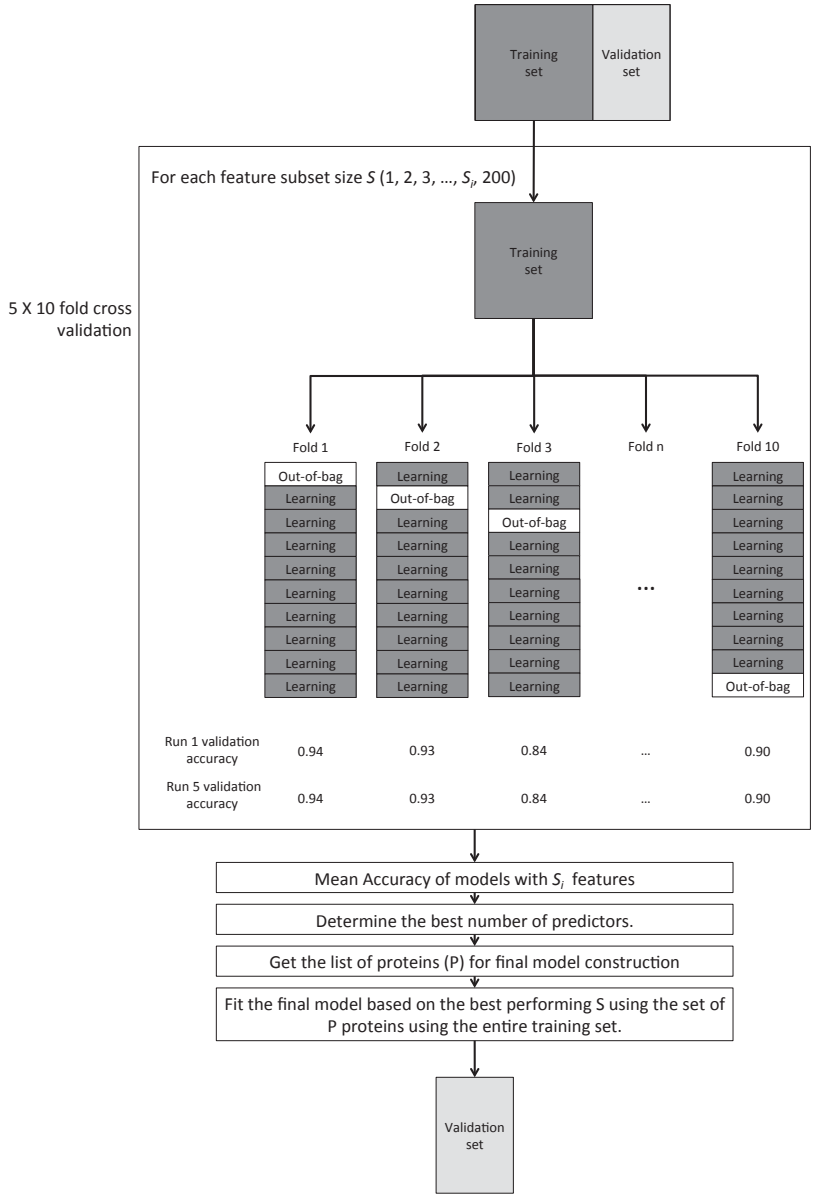
The best performing f may not provide the most parsimonious model. A classification accuracy similar to that of the best performing model might be achieved by using smaller number of features. In this investigation I chose to use the model composed of the smallest set of f_m features that performed with a mean classification accuracy within two percent of the best performing set of f features.

4.1.2.5.4 Summary of Random Forest model building

In brief I used a five x 10-fold cross-validation design using recursive feature elimination to identify the top f features to be used in a final Random Forest Model. Across each of the 50 (5 x 10) cross validation runs I created Random Forest models that ranked features in order of their classification accuracy and then created additional Random Forest models using subsets of those features of sizes $f \in \{1, 2, 3, \cdots, 100, 200\}$. The number of features to be used in the final Random Forest model was the f_m that performed with a mean classification accuracy within two percent of the best performing f across 50 cross validation runs.

The final Random Forest model for each comparison was built using the entire training set using the top f_m features and was tested in the test set. For each comparison I built Random Forests consisting of 500 trees, the number of features sampled at each decision point was equal to the square root of the number of features submitted to the Random Forest algorithm. I used mean classification accuracy as the measure of feature importance.

Figure 4.1.2.1: Random Forest with recursive feature elimination and 5 x 10 fold cross validation.



For each of the 10 pairwise comparisons among the 5 cancer classes the above analysis pipeline was used. First, the dataset was split into a two thirds training set and one third validation set. The majority class was down-sampled to equal the minority class size. 5 x 10 fold cross validation was then used to determine the number of features S and the list of features P to be used to build the final Random Forest model according to mean classification accuracy across the 50 folds. A final Random Forest classifier was constructed using the list of features P using the entire training set. The final classifier was used to predict the classes of the validation set in order to assess the real-world performance of the classifier.

4.1.2.5.5 Measuring model performance

I used the area under the curve (AUC) statistic to measure and compare the performance of each classification model along with model sensitivity, specificity and classification accuracy in the test set. The sensitivity was the proportion of correct positive class classifications and the specificity was the proportion of correct negative class classifications. The positive predictive value (PPV) and negative predictive value (NPV) were not appropriate to measure the discriminative ability of the classification models. The PPV and NPV are influenced by the prevalence of the positive class. As positive class prevalence increases in the test set, PPV increases and NPV decreases. Sensitivity, specificity and accuracy are not influenced by positive class prevalence and were used to measure the discriminative power of the classification models.

4.1.2.5.6 Interpretation of the Random Forest models

The Random Forest algorithm outputs a classification model that consists of a forest of T (500 in this case) trees. The complete description of the forest, is not suitable for human interpretation: Random Forest is a *black box* technique for sample classification. Following Renner et al. (2013) I used a heat map to broadly summarise the Random Forest models. The randomForest implementation in `R` provides summary information about the similarity of the samples using the *proximity* measure. The proximity is a metric measure of the pairwise distance between samples according to the number of times each sample pair is classified by identical decisions across the forest. The *localImp* measure stores information that represents the similarity of the features in the forest according to the importance of each feature for the classification of each sample. I created heat maps of the training set data for each comparison using these two pieces of information. The samples were ordered based on hierarchical clustering of the proximity measure using the Ward method (Ward, 1963). The features were also ordered using hierarchical clustering of the localImp measure using the Ward (Ward, 1963) method. The heat maps were intended to provide some insight in to the construction of the Random Forest models (Figures 4.1.6.2 to 4.1.6.11).

4.1.2.5.7 Gene set analysis of the discriminating features

The Random Forest output included a measure of the importance of each model feature. For each Random Forest model containing more than 10 protein features I investigated which disease pathways, if any, were enriched for the protein features included in the model. I used the WebGestalt (Wang et al., 2013) platform to test for the enrichment of the KEGG Disease Pathways (Kanehisa & Goto, 2000; Kanehisa et al., 2014) with the Random Forest model protein features using Fisher's Exact Test, a False Discovery Rate (FDR) (Benjamini & Hochberg, 1995) q-value <0.1, and using the Entrez gene (Maglott, 2004) protein coding gene set as the background. Pathway gene sets in which more than two genes were covered by the Random Forest model gene list were considered.

The analysis scripts used in this chapter can be found at https://github.com/SutherlandRuss/RS_PhD_scripts.

4.1.3 Results

4.1.3.1 Univariate analyses

The age of individuals differed across the five cancer classes ($F(4, 3044) = 15.22$, $p = 2.49 \times 10^{-12}$) (Table 4.1.3.1). Using t-tests AML samples were found to be significantly younger than samples from all other classes (adenocarcinoma ($t(184.45) = -5.75, p = 3.66 \times 10^{-8}$), glioblastoma ($t(250.02) = -3.42, p = 7.34 \times 10^{-4}$), squamous ($t(204.12) = -5.47, p = 1.3 \times 10^{-7}$), and urothelial ($t(231.8) = -5.94, p = 1.03 \times 10^{-8}$)). The gender distribution was unbalanced across the cancer types ($\chi^2(71) = 35.97$, $p = 2.94 \times 10^{-7}$) (Table 4.1.3.1). All sixteen mutation type frequencies and transition and transversion frequencies differed across the five cancer classes as measured using the Kruskal-Wallis test (Table 4.1.3.2) at the Bonferroni adjusted significance threshold (p<0.0016 : 0.05 / 32 tests stratified by cancer type and stage). When stratified by cancer stages (I,II,III, and IV) there were no differences across the ten mutation type frequencies at the Bonferroni significance threshold (p<0.0016) (Table 4.1.3.2). However, the A>G and T>C transition frequency differed across the cancer stages ($\chi^2(3) = 46.12, p = 5.34 \times 10^{-10}$), as did the A>T and T>A transversion frequency ($\chi^2(3) = 16.18, p = 1.04 \times 10^{-3}$) (Table 4.1.3.2).

Table 4.1.3.1: High-order cancer type demographics

	Total sample	adenocarcinoma	GBM	leukaemia	squamous	urothelial	statistical tests
N	1739	791	275	152	432	89	
mean age (sd)	62.96 (12.67)	64.02 (12.2)	61.21 (12.54)	55.98 (16.39)	63.84 (11.34)	66.66 (11.42)	$F(4, 3044) = 15.22$, $p = 2.49 \times 10^{-12}$
gender N (f/m)	655/1084	337 / 454	101 / 174	72 / 80	119 / 313	26 / 63	$\chi^2(71) = 35.97$, $p = 2.94 \times 10^{-7}$

The age and gender of samples is biased across the cancer classes, as shown by the anova and chisquare results in the statistical tests column.

4.1.3.2 Random Forest prediction models

I used exome sequence derived protein features and clinical features (age, gender and stage where available) using Random Forest with recursive feature elimination and five x 10-fold cross validation to identify sets of proteins that discriminate between 5 cancer classes across 10 pairwise comparisons.

All classification models performed with an AUC above 0.8 (Table 4.1.3.3, and Figure

Table 4.1.3.2: Median frequencies of mutations, and univariate tests across cancer types

		Total sample	adenocarcinoma	glioblastoma	leukaemia	squamous	urothelial	Kruskal-Wallis test across differentiation subtypes
variant frequencies (median / inter-quartile range)	Frame Shift Del	2 (3)	3 (5)	2 (2)	0 (1)	3 (4)	4 (4)	$\chi^2(4)= 289.15$, p=2.38×10^{-61}
	Frame Shift Ins	1 (2)	2 (2)	0 (1)	0 (1)	1 (2)	2 (2)	$\chi^2(4)= 200.52$, p=2.9×10^{-42}
	In Frame Del	0 (1)	1 (1)	0 (1)	0 (0)	1 (2)	1 (2)	$\chi^2(4)= 171.04$, p=6.24×10^{-36}
	In Frame Ins	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	$\chi^2(4)= 19.20$, p=7.17×10^{-4}
	Missense Mutation	55 (80)	49 (50.5)	46 (19)	8 (6)	119 (125.5)	137 (131)	$\chi^2(4)= 704.45$, p=3.79×10^{-151}
	Nonsense Mutation	4 (8)	4 (6)	3 (2)	0 (1)	10 (11)	13 (13)	$\chi^2(4)= 651.56$, p=1.07×10^{-139}
	Nonstop Mutation	0 (0)	0 (0)	0 (0)	0 (0)	0 (1)	0 (1)	$\chi^2(4)= 109.38$, p=9.89×10^{-23}
	RNA	2 (5)	3 (6)	1 (2)	0 (0)	3 (3)	5 (3)	$\chi^2(4)= 390.79$, p=2.72×10^{-83}
	Silent	18 (27)	17 (17)	15 (7)	2 (3)	41 (39.25)	44 (48)	$\chi^2(4)= 698.48$, p=7.45×10^{-150}
	Splice Site	1 (3)	1 (2)	1 (2)	0 (0)	3 (5)	3 (4)	$\chi^2(4)= 379.89$, p=6.17×10^{-81}
Transition / Transversion SNV frequencies (median / inter-quartile range)	A>C_T>G	12 (20)	13 (14)	6 (5)	1 (2)	30 (71.25)	20 (17)	$\chi^2(4)= 735.23$, p=8.21×10^{-158}
	A>G_T>C	40 (48)	29 (38.5)	42 (19)	7 (4)	69 (49.5)	102 (105)	$\chi^2(4)= 689.85$, p=5.51×10^{-148}
	A>T_T>A	5 (9)	6 (8)	3 (3)	0 (1)	11 (19)	4 (4)	$\chi^2(4)= 567.02$, p=2.12×10^{-121}
	C>A_G>T	4 (5)	5 (5)	2 (2)	0 (1)	6 (6)	4 (5)	$\chi^2(4)= 545.86$, p=8.07×10^{-117}
	C>G_G>C	8 (18)	7 (9)	5 (4)	1 (2)	27 (33.25)	47 (71)	$\chi^2(4)= 827.94$, p=6.81×10^{-178}
	C>T_G>A	10 (12)	10 (10)	8 (4.5)	1 (2)	20 (17.25)	14 (9)	$\chi^2(4)= 611.80$, p=4.34×10^{-131}

		stage 1	stage 2	stage 3	stage 4	Kruskal-Wallis test across stages
variant frequencies (median / inter-quartile range)	Frame Shift Del	3 (5)	3 (4)	3 (4)	3 (4)	$\chi^2(3)= 0.18$, p=0.98
	Frame Shift Ins	1 (3)	1 (3)	1 (3)	2 (2)	$\chi^2(3)= 6.06$, p=0.11
	In Frame Del	1 (1)	1 (2)	1 (1)	1 (2)	$\chi^2(3)= 9.74$, p=0.02
	In Frame Ins	0 (0)	0 (0)	0 (0)	0 (0)	$\chi^2(3)= 7.19$, p=0.07
	Missense Mutation	58 (124)	84 (133.75)	65 (120.5)	77 (75.5)	$\chi^2(3)= 11.04$, p=0.01
	Nonsense Mutation	5 (10)	6.5 (12)	6 (11)	6 (8)	$\chi^2(3)= 5.72$, p=0.13
	Nonstop Mutation	0 (0)	0 (1)	0 (0)	0 (0)	$\chi^2(3)= 7.81$, p=0.05
	RNA	3 (5)	2 (5)	3 (5)	3 (5)	$\chi^2(3)= 10.71$, p=0.01
	Silent	20 (39.75)	26.5 (45)	21 (41)	26 (24)	$\chi^2(3)= 9.54$, p=0.02
	Splice Site	2 (3)	1 (4)	2 (4)	2 (3)	$\chi^2(3)= 6.86$, p=0.08
Transition / Transversion SNV frequencies (median / inter-quartile range)	A>C_T>G	16 (65.5)	17 (51.5)	16 (31.5)	16 (13.5)	$\chi^2(3)= 7.38$, p=0.06
	A>G_T>C	34.5 (48.75)	56 (67.5)	43 (61.5)	57 (52.5)	$\chi^2(3)= 46.12$, p=5.34×10^{-10}
	A>T_T>A	7 (13.75)	6 (13)	7 (11)	6 (8)	$\chi^2(3)= 16.16$, p=1.04×10^{-3}
	C>A_G>T	5 (6)	5 (6.25)	5 (6)	5 (4)	$\chi^2(3)= 10.57$, p=0.01
	C>G_G>C	10 (26)	12 (24)	10 (30)	14 (25)	$\chi^2(3)= 3.49$, p=0.32
	C>T_G>A	13 (16)	12.5 (19)	12 (14)	13 (12)	$\chi^2(3)= 1.121$, p=0.77

In almost all cases the frequency of variant types and transition / transversion mutations is biased across both cancer classes (Kruskal-Wallis test across differentiation subtypes column) and across stages for the cancers which had a stage classification (Kruskal-Wallis test across stages column).

4.1.3.1) except for that of the comparison between squamous cell carcinomas and urothelial cell carcinomas (Table 4.1.3.3, and Figure 4.1.3.1). For all classification models, except for the comparison between squamous cell carcinomas and urothelial carcinomas, and between adenocarcinomas and squamous cell carcinomas the classification specificity exceeded the sensitivity. In most models when a sample was classified as the positive class, I could be confident the sample indeed belonged to the positive class.

The model that discriminated between adenocarcinomas and urothelial carcinomas was composed of three features *(*the genes *APC, VHL*, and cancer stage), and performed with a high sensitivity of 0.967 (Table 4.1.3.3). The squamous cell carcinoma and urothelial carcinoma classes were the most difficult to distinguish as shown by the AUC of 0.771 (Table 4.1.3.3, and Figure 4.1.3.1). One hundred features were included in the model comparing the squamous cell carcinomas and urothelial carcinomas: together with the lowest AUC, this may indicate that overfitting had occurred. In addition, mean classification accuracies that were negative were recorded for the six least important features *CHD7*, *MLL3*, *DNAH17*, *OR6T1*, *MTMR2*, and *PLCE1*.

For each comparison the model features ranked according to their mean classification accuracy in the training set are shown in Table 4.1.3.4. *APC*, and *VHL* were part of all models that included adenocarcinomas. *TP53* was part of all models that compared squamous cell carcinomas. *PTEN*, and *EGFR* were included in all glioblastoma models, and *TTN* was part of every AML model. There were no proteins that distinguished urothelial carcinomas from all other cancer types.

By inspecting the heatmaps generated from the Random Forest training data, some insight can be gained in to how the features in the models separated the classes in each comparison. When comparing adenocarcinomas and squamous cell carcinomas, *VHL* and *APC* were mutated only in adenocarcinomas, and *KRAS* mutations appeared to be only in the squamous cell carcinomas (Figure 4.1.6.2 on page 128). Among the other features included in the model, more complex interactions may explain their classification. For the adenocarcinoma and urothelial carcinoma comparison again *VHL* and *APC* were mutated

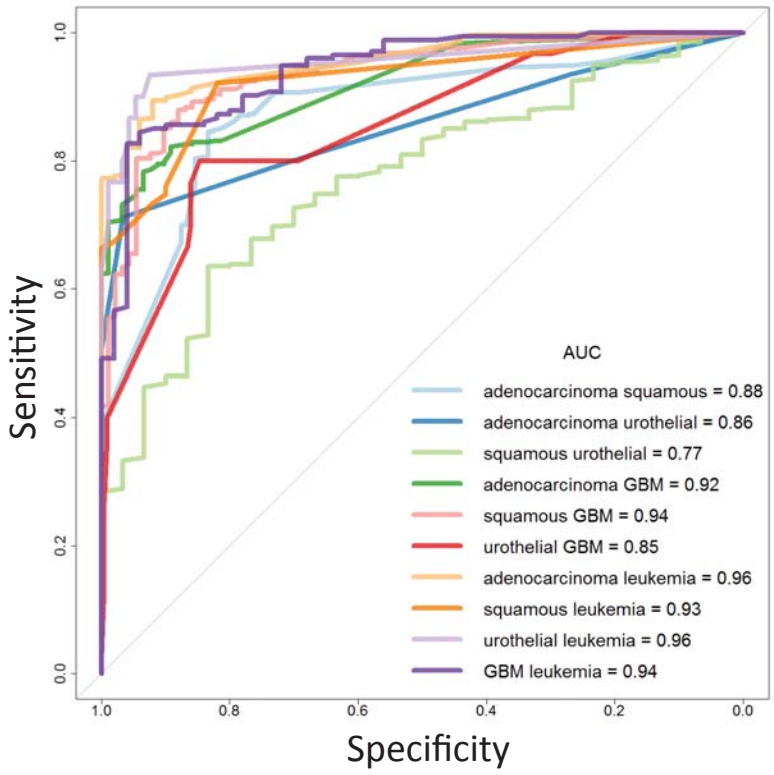
only among adenocarcinomas, and the stage I cancers were all adenocarcinomas (Figure 129). When comparing urothelial carcinomas and glioblastomas *ATM*, and *MLL2,* were mutated only in urothelial carcinomas, and TTN, *ARID1A*, and *MLL3* mutations were most common among the urothelial carcinomas. However, *EGFR* mutations were exclusive to glioblastomas, and PTEN mutations were almost exclusive to glioblastoma (Figure 4.1.6.7 on page 133). For the comparison between squamous cell carcinomas and AML all of the genes were most frequently mutated in the squamous cell carcinomas, with AML mutations only occurring in *TTN*, *TP53*, and *CDKN2A* (Figure 4.1.6.9 on page 135). Across all of the AML comparisons it appeared that the mutated genes that best characterised AML were *RUNX1*, *DNMT3A*, *NPM1*, and *FLT3*.

Table 4.1.3.3: Random Forest classification model statistics

Positive class	Negative class	Sensitivity	Specificity	Accuracy	AUC
adenocarcinoma	**squamous**	0.85	0.83	0.84	0.88
adenocarcinoma	**urothelial**	0.71	0.97	0.72	0.86
squamous	**urothelial**	0.79	0.53	0.77	0.77
adenocarcinoma	**GBM**	0.78	0.94	0.80	0.92
squamous	**GBM**	0.78	0.95	0.83	0.94
urothelial	**GBM**	0.80	0.83	0.83	0.85
adenocarcinoma	**leukemia**	0.81	0.96	0.82	0.96
squamous	**leukemia**	0.76	0.90	0.78	0.93
urothelial	**leukemia**	0.77	0.97	0.92	0.96
GBM	**leukemia**	0.86	0.86	0.86	0.94

The classification statistics for each pairwise Random Forest comparison are shown in each row. **Positive class** = the cancer type that was considered a true positive when successfully classified for the purpose of calculating the classification statistics. **Negative class** = the cancer type that was considered a true negative when successfully classified for the purpose of calculating the classification statistics.

Figure 4.1.3.1: Random Forest classification model ROC curves



The ROC curves for all pairwise Random Forest comparisons show that an area under the curve (AUC) of at least 0.80 was achieved in all comparisons except for that between squamous cell carcinomas and urothelial carcinomas.

4.1.3.2.1 Random Forest with recursive feature elimination

For each of the comparisons I aimed to identify the model that best discriminated between the two cancer classes and simultaneously the set of features used in the model. Classification performance in the training set can increase even if uninformative features are added to the model; the classifier can begin to model the noise in the training set, leading to over-fitting. I attempted to avoid over-fitting by choosing the models with the smallest number of features and a mean classification accuracy within two percent of the best performing model. The extent of model refinement using this accuracy tolerance is visualised in supplementary figures 4.1.6.1 on page 126. In most cases the best performing models may be over-trained, containing more features than models of close to equal mean classification accuracy (Figure 4.1.6.1). For three of the pairwise comparisons, ([adenocarcinoma, squamous], [squamous, GBM], and [squamous, AML]) the best performing Random Forest model contained the complete set of features that were input to the model.

4.1.3.3 KEGG disease pathway enrichment

KEGG disease pathway enrichment analysis was conducted for the models discriminating squamous cell carcinomas from urothelial cell carcinomas, adenocarcinomas from glioblastoma, squamous cell carcinoma from glioblastoma, adenocarcinomas from AML, urothelial cell carcinomas from AML and glioblastoma from AML. These models were all composed of more than 10 features. For the squamous cell carcinoma and glioblastoma comparison, enrichment analysis using KEGG disease pathways (Kanehisa & Goto, 2000; Kanehisa et al., 2014) identified 14 disease pathways enriched for the 53 proteins. The top 10 enriched pathways were all cancers, or cancer related, with the most highly enriched pathway being the 'p53 signalling pathway' (adjusted p=3.00×10^{-4}). However, none of the enriched pathways were squamous cell carcinomas. For the Adenocarcinoma and AML comparison, the KEGG disease pathway enrichment analysis identified six cancer pathways enriched for genes in the model, AML being the most highly enriched pathway (FDR adjusted p= 2.00×10^{-4}). For the AML and Glioblastoma comparison KEGG disease pathway enrichment analysis returned 49

Table 4.1.3.4: Random Forest model features and variable importance measures

adenocarcinoma, squamous			adenocarcinoma, urothelial			squamous, urothelial										adenocarcinoma, glioblastoma		
	Features	Importance		Features	Importance		Features	Importance		Features	Importance		Features	Importance			Features	Importance
1	*TP53*	39.30	1	*APC*	24.83	1	*TP53*	16.81	31	*ZNF845*	4.17	61	*MYH8*	2.62	91	*ZNF513*	0.55	1	*VHL*	60.95
2	*APC*	32.24	2	*stage*	24.46	2	*stage*	11.31	32	*PRX*	4.11	62	*DCHS2*	2.62	92	*CASP8*	0.53	2	*APC*	45.52
3	*KRAS*	28.11	3	*VHL*	23.64	3	*CDKN2A*	9.20	33	*FAT1*	4.00	63	*ANKRD28*	2.62	93	*PTPN14*	0.33	3	*PBRM1*	40.78
4	*CSMD3*	21.56				4	*SLITRK3*	8.50	34	*FNDC3A*	3.98	64	*HNRPDL*	2.51	94	*PRG4*	0.18	4	*KRAS*	26.34
5	*VHL*	19.72				5	*CSMD3*	8.33	35	*CNTLN*	3.93	65	*OR2J3*	2.44	95	*CHD7*	-0.04	5	*PTEN*	23.44
6	*NOTCH1*	17.94				6	*SI*	7.92	36	*HTR1E*	3.92	66	*KIAA2018*	2.41	96	*MLL3*	-0.52	6	*RYR1*	20.25
7	*CASP8*	16.19				7	*ARID1A*	7.75	37	*PCDH20*	3.80	67	*INO80*	2.36	97	*DNAH17*	-0.67	7	*CSMD1*	19.37
8	*PIK3CA*	15.84				8	*KLHL1*	6.31	38	*SRP68*	3.70	68	*ZNF761*	2.34	98	*OR6T1*	-0.68	8	*EGFR*	18.93
						9	*PRAMEF11*	6.19	39	*ELF3*	3.67	69	*PCNT*	2.32	99	*MTMR2*	-0.89	9	*CDH10*	18.26
						10	*MYH7*	6.11	40	*AVPR1B*	3.66	70	*NCAPD2*	2.30	100	*PLCE1*	-1.14	10	*SETD2*	16.65
						11	*ERCC6L2*	6.01	41	*CYP3A7*	3.49	71	*FLG2*	2.27				11	*CROCCP2*	16.48
						12	*RXRA*	5.73	42	*SLIT2*	3.47	72	*AJUBA*	2.26				12	*SLC22A6*	16.32
						13	*COL3A1*	5.71	43	*ARHGAP32*	3.41	73	*WWC1*	2.25				13	*TNN*	15.28
						14	*HLA-B*	5.66	44	*C20orf26*	3.33	74	*DHX34*	2.18				14	*VPS13B*	14.27
						15	*CDH19*	5.57	45	*ATP13A2*	3.33	75	*PAPOLA*	2.15				15	*GRID1*	13.45
						16	*OR4A15*	5.43	46	*HERC1*	3.32	76	*ARVCF*	2.10				16	*ZFPM2*	13.04
						17	*ANK3*	5.11	47	*CDH9*	3.25	77	*ZNF91*	1.93				17	*MMP2*	12.83
						18	*FAM135B*	5.10	48	*PPP2R3A*	3.24	78	*CCDC40*	1.91				18	*SMAD4*	12.24
						19	*ADAMTS20*	5.09	49	*ACACA*	3.16	79	*OR5L1*	1.69				19	*TNC*	11.26
						20	*ANO5*	5.01	50	*ADAMTS7*	3.16	80	age	1.56				20	*RP1L1*	10.98
						21	*DOT1L*	5.00	51	*TICRR*	3.15	81	*COL7A1*	1.52				21	MYO16	10.83
						22	*ZNF626*	4.92	52	*ANKS1B*	3.12	82	*OTUD4*	1.47				22	IPO7	10.76
						23	*MLL*	4.72	53	*EGFR*	3.05	83	*ASPM*	1.43				23	PIK3R1	10.45
						24	*ERCC2*	4.70	54	*LOC643733*	2.95	84	*KIF24*	1.39				24	BAP1	10.07
						25	*TACC3*	4.70	55	*GAK*	2.86	85	*CSMD1*	1.34				25	FAT3	9.89
						26	*ZFHX3*	4.59	56	*A2M*	2.79	86	*VRK2*	1.33				26	MTOR	9.68
						27	*SPTA1*	4.55	57	*PADI2*	2.74	87	*EP300*	1.30				27	PCDH20	9.04
						28	*KIAA0754*	4.47	58	*RNF40*	2.71	88	*GRM8*	1.10						
						29	*CDKN1A*	4.25	59	*BBS9*	2.65	89	*DNAH1*	0.94						
						30	*CACNA1H*	4.18	60	*UBBP4*	2.63	90	*ANK2*	0.91						

Features = The features included in each Random Forest model. **Importance** = The mean classification accuracy of each feature across the cross-validation folds.

Table 4.1.3.4: Random Forest Model Features and variable importance measures continued.

	squamous, glioblastoma				urothelial, GBM		adenocarcinoma, leukemia		squamous, leukemia		urothelial, leukemia		GBM, leukemia	
	Features	Importance	Features	Importance	Features	Importance	Features	Importance	Features	Importance	Features	Importance	Features	Importance
1	*MLL2*	29.12	*ZFHX4*	8.45	*MLL2*	21.50	*VHL*	39.67	*TTN*	32.78	*TTN*	30.05	*PTEN*	29.72
2	*CDKN2A*	24.83	*PAPPA2*	8.02	*PTEN*	20.91	*APC*	34.84	*LRP1B*	26.77	*MLL3*	20.04	*EGFR*	29.16
3	*CSMD3*	24.69	*DYNC1H1*	7.58	*TTN*	20.62	*PBRM1*	21.99	*SYNE1*	25.41	*MLL2*	17.61	*DNMT3A*	21.32
4	*FAT1*	22.70	*SEMA5A*	7.44	*EGFR*	20.24	*MUC16*	19.39	*NOTCH1*	21.86	*MUC16*	15.05	*PCLO*	16.16
5	*NOTCH1*	22.16	*XIRP2*	7.43	*ARID1A*	16.39	*TTN*	19.12	*TP53*	21.47	*FRG1B*	14.68	*MUC16*	16.04
6	*TP53*	21.38	*FN1*	7.41	*MLL3*	15.03	*LRP1B*	17.79	*CDKN2A*	19.74	*RYR2*	14.32	*FRAS1*	15.38
7	*PTEN*	19.18	*JMJD1C*	7.38	*ATM*	13.49	*FLT3*	17.13			*ARID1A*	13.97	*FLT3*	14.69
8	*CASP8*	14.05	*SLC4A4*	7.36			*DNMT3A*	16.00			*MUC17*	13.21	*NPM1*	13.76
9	*EGFR*	13.29	*ASXL3*	6.99			*MUC4*	13.00			*COL6A6*	12.82	*RUNX1*	13.00
10	*LRP1B*	12.94	*POM121L9P*	6.99			*IGFN1*	12.86			*MACF1*	12.20	*PIK3CA*	12.76
11	*PKHD1L1*	12.90	*FAM135B*	6.84			*RUNX1*	12.76			*SYNE1*	11.94	*TUBBP5*	12.31
12	*PTPRT*	12.82	*SMARCA4*	6.82			*PIK3CA*	12.60			*DNMT3A*	11.62	*NRAS*	12.26
13	*ZNF804B*	12.38	*KEL*	6.76			*RYR2*	12.17			*ATM*	11.31	*PIK3R1*	11.57
14	*COL11A1*	11.87	*VCAN*	5.60			*BAP1*	11.55			*RUNX1*	10.49	*FLG*	10.42
15	*CUBN*	11.83	*PXDN*	5.54			*XIRP2*	11.55			*FLG*	8.66	*TTN*	9.90
16	*CDH10*	11.54	*SI*	5.52			*age*	11.46					*age*	8.53
17	*ERBB4*	11.29	*GRID2*	5.47			*PLG*	11.43					*NF1*	8.47
18	*NFE2L2*	10.33	*TTN*	5.38			*NPM1*	10.78					*MUC17*	8.03
19	*NAV3*	10.28	*ANKRD30A*	5.05			*VPS13A*	10.74					*TP53*	7.87
20	*ENSG00000161103*	9.95	*TRHDE*	4.65			*AHNAK2*	10.19					*PKHD1*	7.37
21	*KIAA1109*	9.86	*OR4M1*	4.35			*USH2A*	9.73					*USH2A*	6.74
22	*LRRK2*	9.51	*ASPM*	2.52			*HMCN1*	9.13					*HMCN1*	6.15
23	*MALAT1*	9.30	*NBPF10*	1.80			*SETD2*	8.88					*SETD2*	6.08
24	*SYNE1*	9.30					*RP1L1*	8.67					*RP1L1*	5.90
25	*ADCY8*	8.99					*SYNE1*	8.18						
26	*FAT4*	8.94					*TRIOBP*	6.99						
27	*FBXW7*	8.73					*OPCML*	6.32						
28	*EPHA2*	8.68												
29	*CSMD1*	8.63												
30	*LDHD*	8.60												

Features = The features included in each Random Forest model. **Importance** = The mean classification accuracy of each feature across the cross-validation folds. The variables included in each Random Forest model were ranked according to their mean classification accuracy in the training set.

enriched pathways including Acute Myeloid Leukaemia (FDR adjusted p=3.67×10^{-8}) and Glioma (adjusted p=2.23×10^{-9}). The Random Forest models that discriminated between the cancer types were enriched in gene sets that either directly correspond to the appropriate cancers, or to cancer related processes.

4.1.4 Discussion

I have identified sets of protein features that discriminate between each of the five high-order cancer types present in the Pan-Cancer data set. I found that many of the features included in the models were already known to be either significantly mutated in a subset of the Pan-Cancer cancers, or exclusively mutated in one Pan-Cancer cancer type (e.g. *VHL* in *KIRC*, *APC* in COADREAD, and *DNMT3A* in AML). This indicated that the features that were most important for discriminating between the cancer types were those which were significantly differentially mutated across the cancer types at the univariate level.

I found that *CASP8* was important for discrimination between adenocarcinomas and squamous cell carcinomas, and squamous cell carcinomas and glioblastoma. This relationship was not identified in early Pan-Cancer analysis papers, nor was *CASP8* found to be significantly mutated (Kandoth et al., 2013a), but it has subsequently been associated with head and neck squamous cell carcinoma (Hoadley et al., 2014; Lawrence et al., 2015). *FRG1B* was also implicated in head and neck squamous cell carcinoma by Hoadley et al. (2014) and Lawrence et al. (2015), in contrast to this analysis where it was important for the discrimination of urothelial carcinomas and AML. *NOTCH1* was not significantly mutated in adenocarcinomas (Kandoth et al., 2013a). However, Lawrence et al. (2015) found *NOTCH1* inactivating mutations in 19 percent of head and neck squamous cell carcinomas, validating the inclusion of *NOTCH1* in the model to discriminate between adenocarcinomas and squamous cell carcinomas.

For the adenocarcinoma and glioblastoma comparison *PBRM1* mutations were known to be specific to kidney renal clear cell carcinoma (Kandoth et al., 2013a). *PTEN* and *EGFR* were found to be most frequently mutated in glioblastoma samples (Kandoth et al., 2013a),

and appeared in every model that included glioblastoma.

For the urothelial and glioblastoma comparison, *MLL2* was more frequently mutated in bladder urothelial cancer than in glioblastoma (Kandoth et al., 2013a; Hoadley et al., 2014). *PTEN* and *EGFR*, frequently mutated in glioblastoma (Kandoth et al., 2013a) were also included in the model. *ARID1A,* and *ATM* were significantly mutated in TCGA bladder urothelial carcinoma data (Weinstein et al., 2014).

For the squamous cell and urothelial carcinoma comparison *CDKN2A* was more frequently mutated in head and neck squamous cell carcinoma and lung squamous cell carcinoma in comparison to bladder urothelial carcinoma (Kandoth et al., 2013a; Lawrence et al., 2015). When comparing bladder urothelial carcinoma and AML the genes *MLL3*, *and MLL2* were included in the model, both of which were mutated at higher frequency in bladder urothelial carcinoma than in AML (Kandoth et al., 2013a; Hoadley et al., 2014).

For the comparison between glioblastoma and leukaemia, *DNMT3A* has been found to be mutated in a higher percentage of leukaemia samples than glioblastoma samples (Kandoth et al., 2013a). Mutations in the genes *DNMT3A*, *RUNX1*, *NPM1*, and *FLT3* that best discriminated AML from other cancer types also define three AML subtypes (Kandoth et al., 2013a).

The Random Forest models include genes that in other cancer genome analyses have potentially been spuriously called as significantly functionally mutated, most notably *CSMD3*, *CSMD1*, *LRP1B*, *RYR2*, *RYR1* and *PCLO* (Lawrence et al., 2013). Those genes may have been incorrectly labelled as significantly mutated in other analyses because of the incorrect assumptions of the background mutation rate of these genes by tools that assume a uniform genome-wide background mutation rate. Many of the genes are late replicating, they have a high local background mutation rate, and are expressed at low levels (Lawrence et al., 2013). It is possible that differences in the overall mutation frequency of each cancer have caused these spurious genes to be included in the models as being biologically discriminative because they indicate a difference between the local background mutation rate across the cancer types. In a further analysis it may be useful to use MutSigCV (Lawrence et al., 2013)

gene scores for mutation significance to weight the binary mutation matrix. In this way the effect of a local increase in background mutation rate at these genes may be accounted for and biologically discriminative genes may be included in the models.

Further evidence for background mutation effects may come from AML comparisons. The gene *TTN* was the only gene included in models for all AML comparisons. *TTN* is the longest human gene, and AML had the lowest mutation frequency of all of the Pan-Cancer cancers (Kandoth et al., 2013a). *TTN* was rarely mutated in AML, and higher background mutation rates among other cancers may be indicated as a mutation in the largest gene, *TTN*, which is likely to accumulate passenger mutations. Indeed, using the Wilcoxon rank sum test samples carrying TTN functional mutations had a higher overall SNV mutation frequency (Mdn=174) than samples in which TTN was not mutated (Mdn=63) ($W = 123350, p < 2.2 \times 10^{-16}$).

The set of cancers compared in this analysis is not comprehensive. It is reasonable to assume that the entire set of genes that discriminates between each of the five cancer classes in the Pan-Cancer dataset has not been identified. For example, *PTEN* mutations were important for discriminating glioblastoma from the other cancer types. However, if endometrial carcinoma had been included in the analysis *PTEN* may not have been included in the model discriminating glioblastoma from adenocarcinoma, because *PTEN* mutations occur in endometrioid endometrial adenocarcinomas (Kandoth et al., 2013b).

4.1.5 Conclusions

By using Random Forest analysis I created models that showed good ability to discriminate between the five high-order types in a pairwise manner. For most of the pairwise comparisons the proteins that were found to be most important for cancer type discrimination were those which either, were exclusively mutated in a single cancer type, or were more commonly mutated in a single cancer type. For example *VHL* and *APC* were always included in models to discriminate adenocarcinomas from other cancer types because they were exclusively mutated in kidney renal clear cell carcinoma (Kandoth et al., 2013a) and colorectal adenocarcinoma respectively.

The discriminative models included proteins which in other studies have been found to be spuriously labelled as significantly mutated, including: CSMD3 in ovarian carcinoma (Lawrence et al., 2013); and *LRP1B*, *MUC16*, *MUC4*, *CSMD1*, *PCLO*, and *RYR2* in lung squamous cell carcinoma (Lawrence et al., 2013). These features should be treated with caution. The reason they discriminate between cancer types may not be because they are discriminative driver genes that are important for cancer progression, but because they are susceptible to the accumulation of passenger mutations associated with late replication time (in the case of *CSMD3*) and low levels of expression (Lawrence et al., 2013; Stamatoyannopoulos et al., 2009; Pleasance et al., 2010).

The Random Forest approach to identifying features that discriminate between cancer types may be useful to create tools to classify cancers of unknown primary origin. By providing a prediction of the tissue of origin for a metastatic cancer of unknown primary origin, clinicians may use the information to guide patient treatment; an idea which is developed in the next section of this chapter.

4.1.6 Supplementary Materials

Table 4.1.6.1: cancer stage distribution across cancers

Stage	BLCA	BRCA	COAD	GBM	HNSC	KIRC	LAML	LUAD	LUSC	OV	READ	UCEC
Not recorded	0	0	0	275	0	0	152	0	0	0	0	0
1	1	126	30	0	17	197	0	80	95	0	18	164
2	21	436	57	0	46	40	0	31	36	14	25	20
3	35	172	44	0	42	113	0	35	38	247	16	50
4	32	15	21	0	155	67	0	9	3	53	8	13

The contingency table shows the counts of cancers by stage (rows) and cancer type (columns). **BLCA** = bladder, **BRCA** = breast, **COAD** = colon adenocarcinoma, **GBM**, = glioblastoma, **HNSC** = head and neck squamous cell carcinoma, **KIRC** = kidney, **LAML** = acute myeloid leukaemia, **LUAD** = lung adenocarcinoma, **LUSC** = lung squamous cell carcinoma, **OV** = ovarian cystadenocarcinoma, **READ** = rectum adenocarcinoma, **UCEC** = endometrial carcinoma.

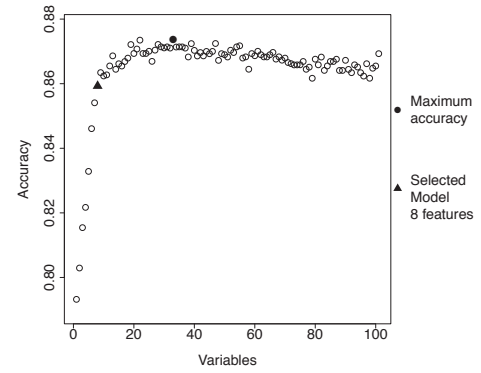
Table 4.1.6.2: cancer stage distribution across high-order cancer types

Stage	adenocarcinoma	glioblastoma	AML	squamous	urothelial
Not recorded	0	275	152	0	0
1	615	0	0	112	1
2	623	0	0	82	21
3	677	0	0	80	35
4	186	0	0	158	32

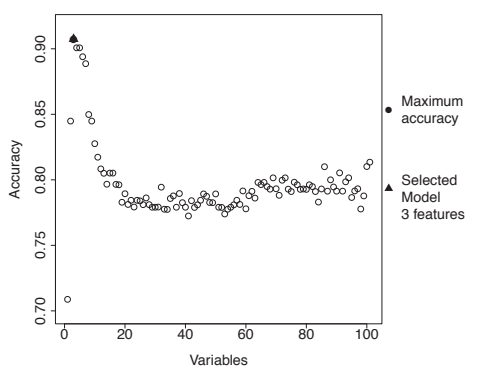
The contingency table shows the counts of cancers by stage (rows) and high-order cancer type (columns). **Adenocarcinoma** = breast, colon, rectum, ovarian, kidney, lung adenocarcinoma; **AML** = acute myeloid leukaemia; **squamous** = lung squamous cell, head and neck squamous cell; **urothelial** = bladder urothelial carcinoma.

Figure 4.1.6.1: Pairwise Random Forest classification accuracy plots

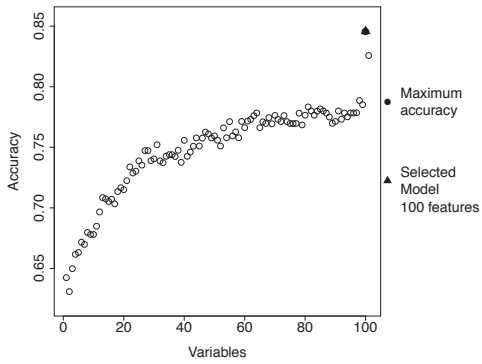
(a) adenocarcinoma / squamous cell carcinoma



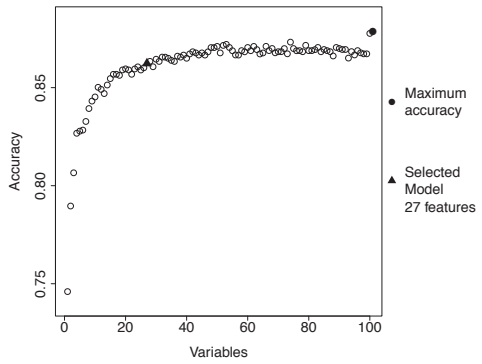
(b) adenocarcinoma / urothelial



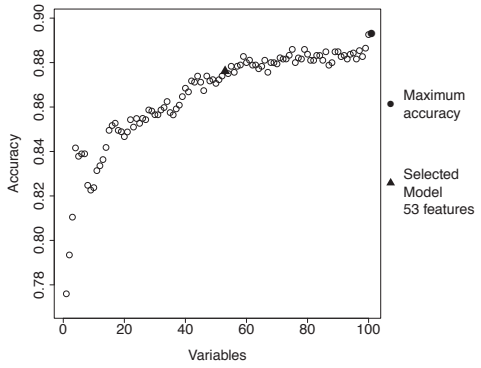
(c) squamous / urothelial



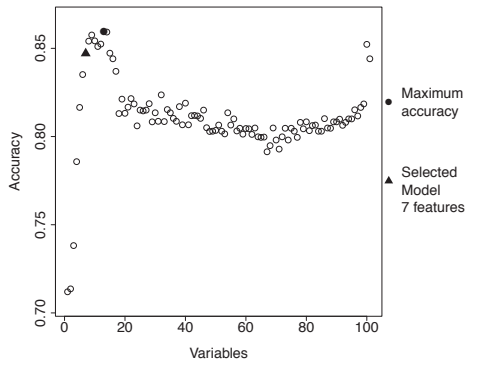
(d) adenocarcinoma / glioblastoma



(e) squamous / glioblastoma



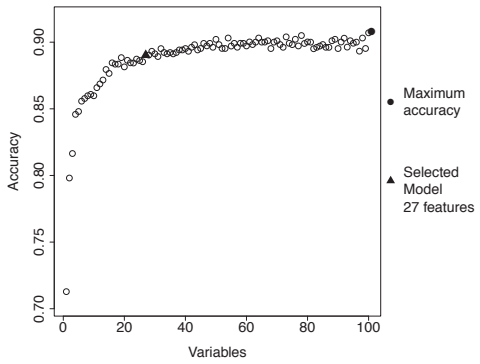
(f) urothelial / glioblastoma



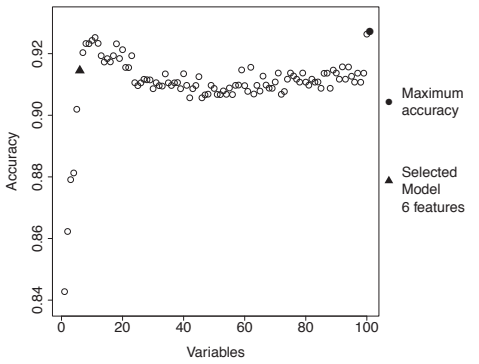
For each comparison the mean classification accuracy achieved (x-axis), by Random Forest models of feature set sizes ranging from 1-100, 200. The 'selected models', contained the smallest number of features that performed with classification accuracy that was at most 2% lower than the model with maximum accuracy. For figures (4.1.6.1b, and 4.1.6.1f), the models built to compare adenocarcinomas with urothelial carcinomas, and urothelial carcinomas with glioblastoma were overfitting the training data when using more than around 20 features. The model selection process selected models that used small numbers of features and performed with high accuracy.

Figure 4.1.6.1: classification accuracy plots continued.

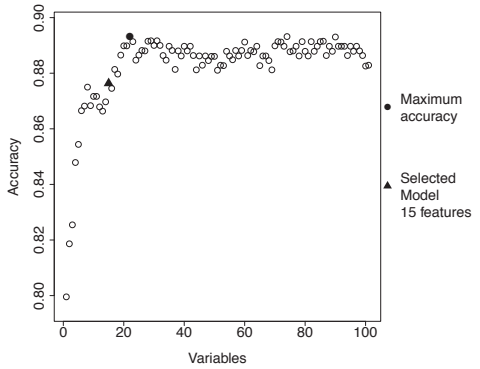
(g) adenocarcinoma / AML



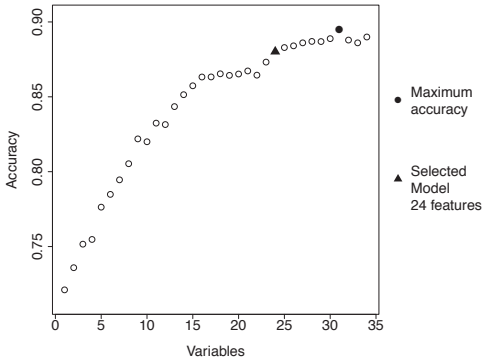
(h) squamous / AML



(i) urothelial / AML

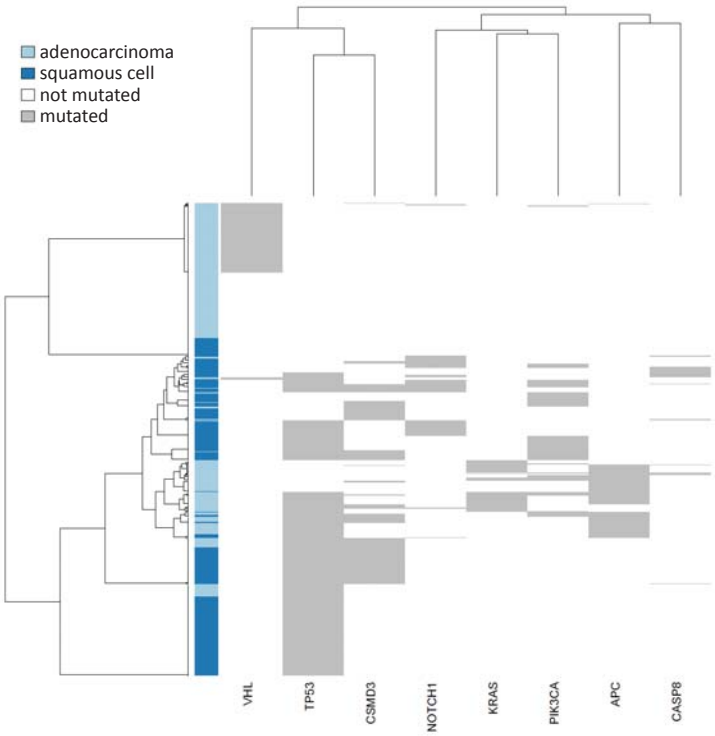


(j) glioblastoma / AML



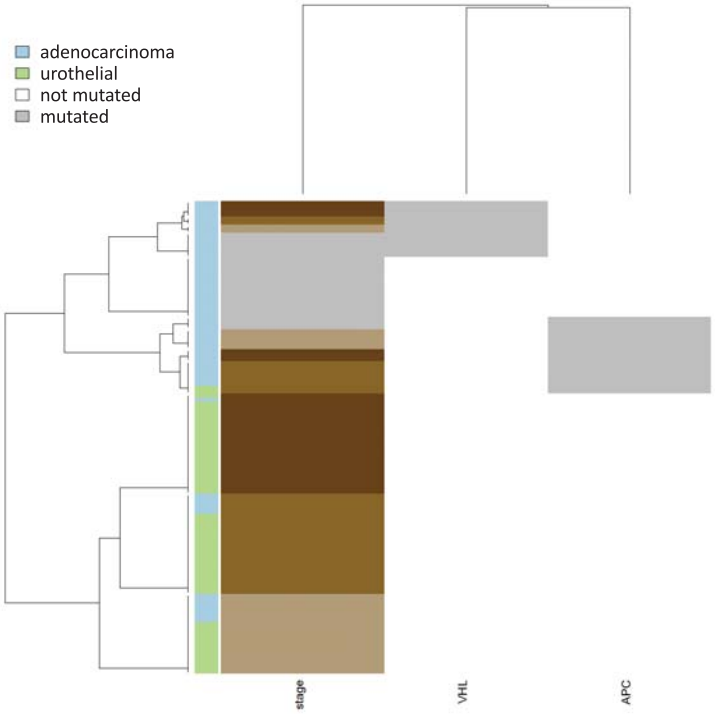
For each comparison the mean classification accuracy achieved (x-axis), by Random Forest models of feature set sizes ranging from 1-100, 200. The 'selected models', contained the smallest number of features that performed with classification accuracy that was at most 2% lower than the model with maximum accuracy. The squamous cell carcinoma and AML comparison (Figure 4.1.6.1h) demonstrates how the model selection process has moved from a model containing 200 features that performed with a mean accuracy of 0.93, to a more parsimonious model of 6 features that performed with a mean accuracy of 0.91.

Figure 4.1.6.2: Adenocarcinoma / squamous cell carcinoma heat map



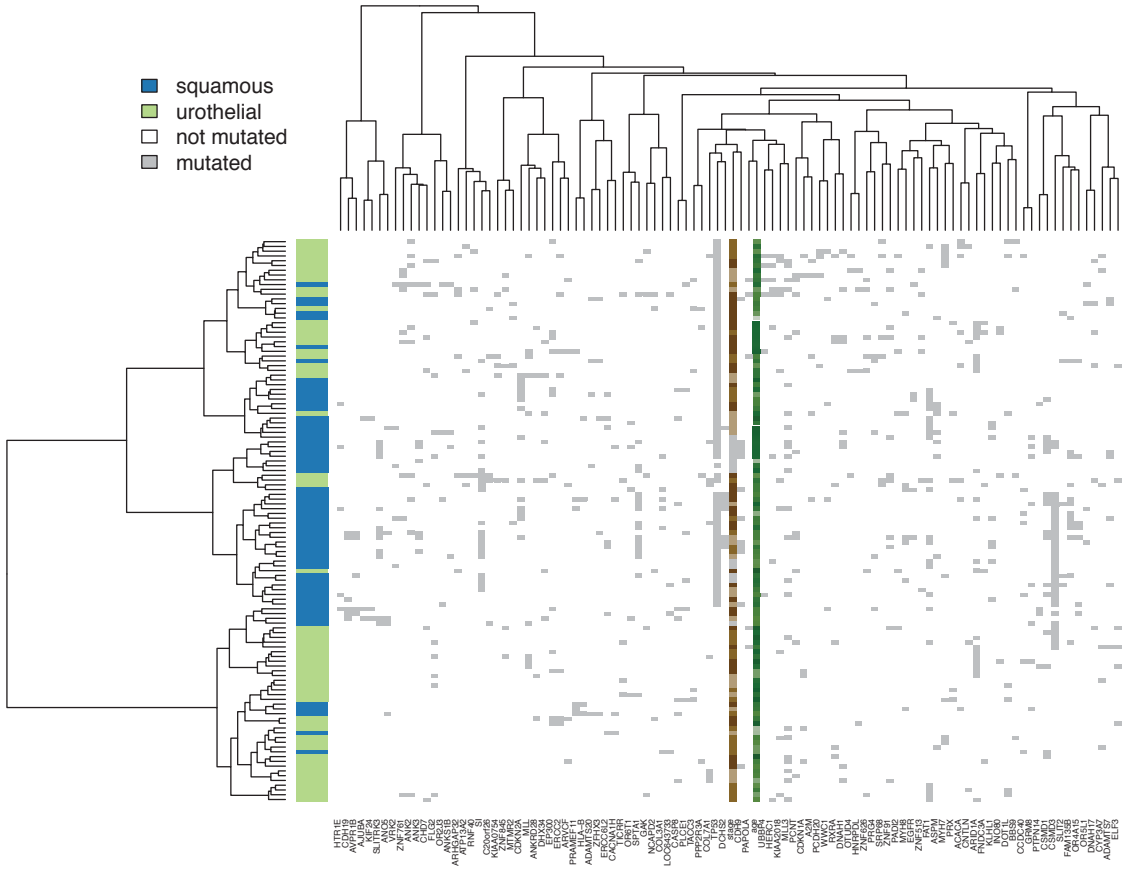
The rows represent samples and columns represent genes in this heatmap. The coloured column to the left of the heatmap indicates the class to which each sample belongs. The grey markers indicate in which gene each sample carried a functional mutation.

Figure 4.1.6.3: Adenocarcinoma / urothelial carcinoma heat map



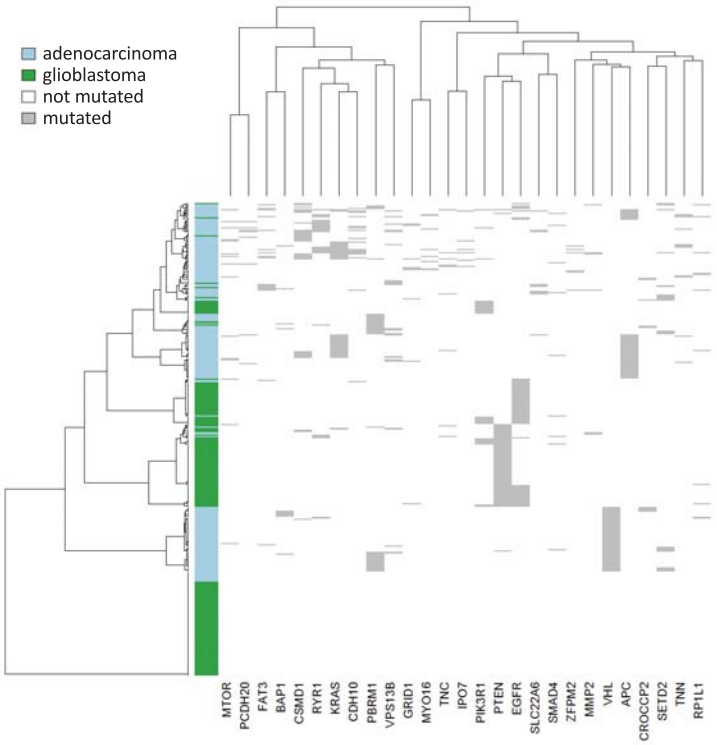
The rows represent samples and columns represent genes in this heatmap. The coloured column to the left of the heatmap indicates the class to which each sample belongs. The grey markers indicate in which gene each sample carried a functional mutation. The brown scaled column indicates cancer stage, progressing from stage 1 (grey) to stage 4 (dark brown)

Figure 4.1.6.4: Squamous cell carcinoma / urothelial carcinoma heat map



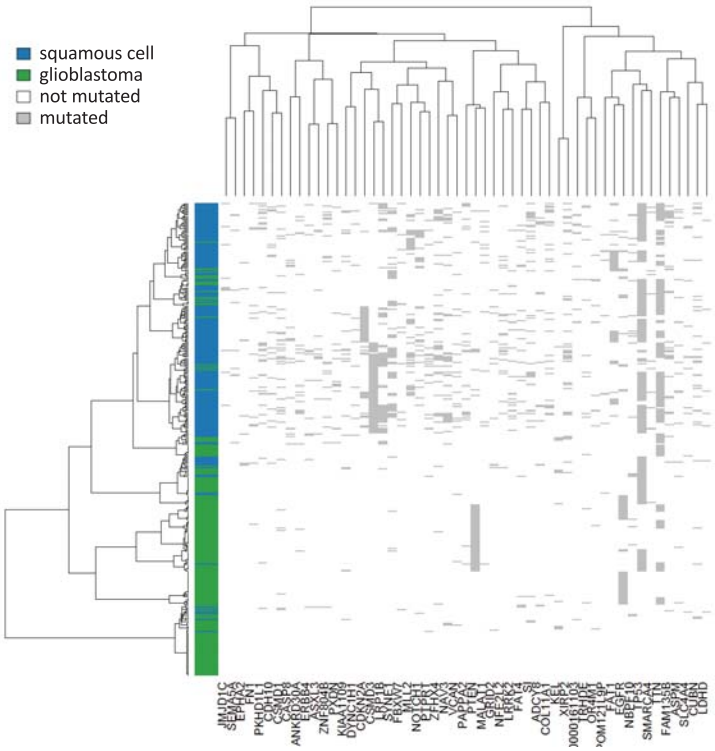
The rows represent samples and columns represent genes in this heatmap. The coloured column to the left of the heatmap indicates the class to which each sample belongs. The grey markers indicate in which gene each sample carried a functional mutation. The brown scaled column indicates cancer stage, progressing from stage 1 (grey) to stage 4 (dark brown). The green column indicates sample age ranging from grey to dark green with increasing age.

Figure 4.1.6.5: Adenocarcinoma / glioblastoma heat map



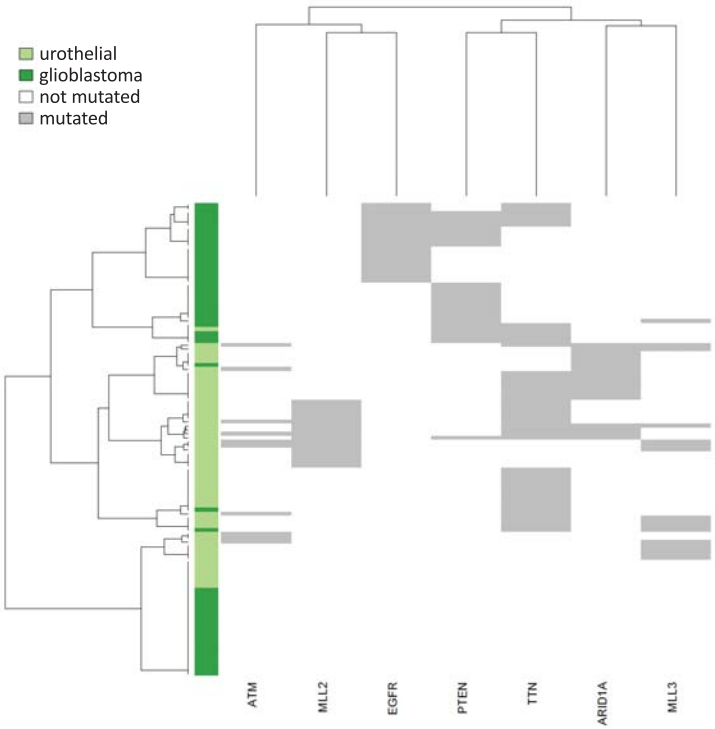
The rows represent samples and columns represent genes in this heatmap. The coloured column to the left of the heatmap indicates the class to which each sample belongs. The grey markers indicate in which gene each sample carried a functional mutation.

Figure 4.1.6.6: Squamous cell carcinoma / glioblastoma heat map



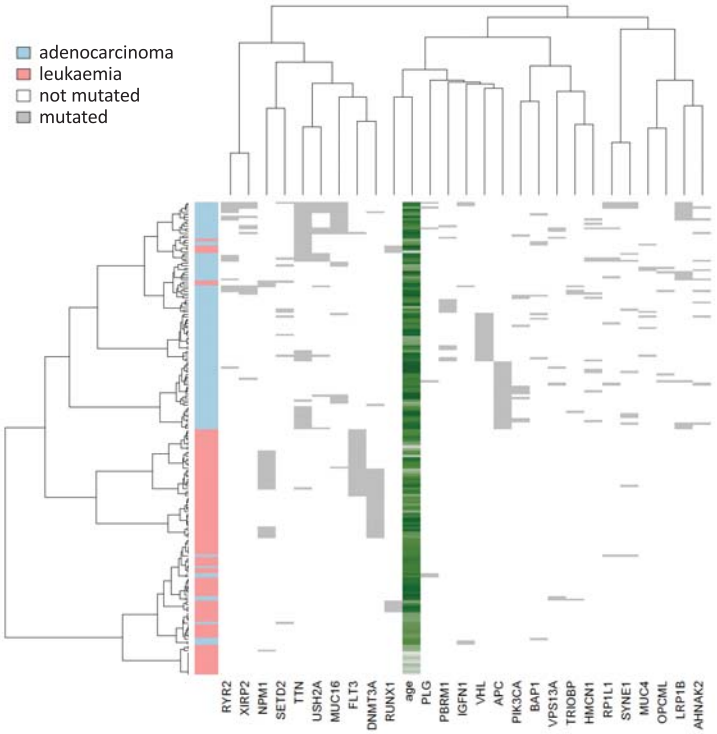
The rows represent samples and columns represent genes in this heatmap. The coloured column to the left of the heatmap indicates the class to which each sample belongs. The grey markers indicate in which gene each sample carried a functional mutation.

Figure 4.1.6.7: Urothelial / glioblastoma heat map



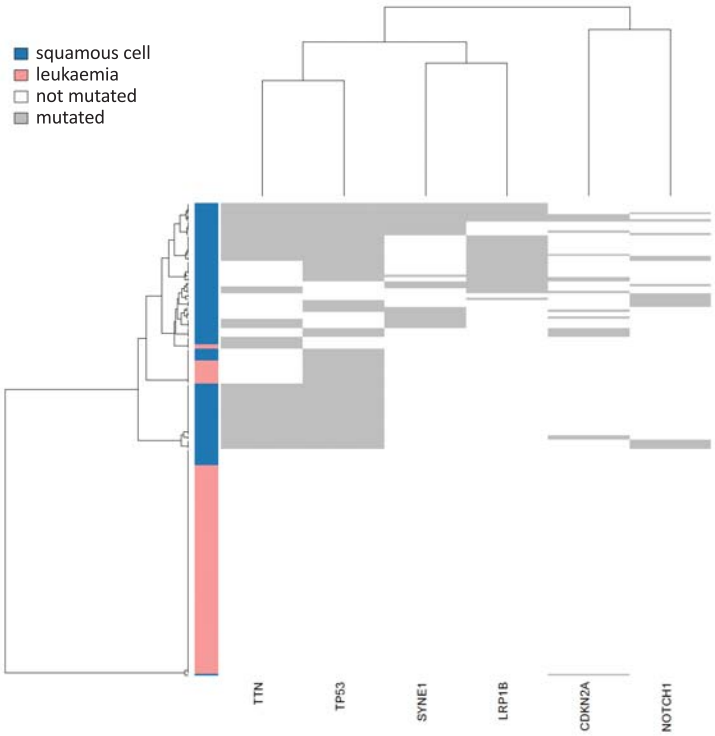
The rows represent samples and columns represent genes in this heatmap. The coloured column to the left of the heatmap indicates the class to which each sample belongs. The grey markers indicate in which gene each sample carried a functional mutation.

Figure 4.1.6.8: Adenocarcinoma / leukemia heat map



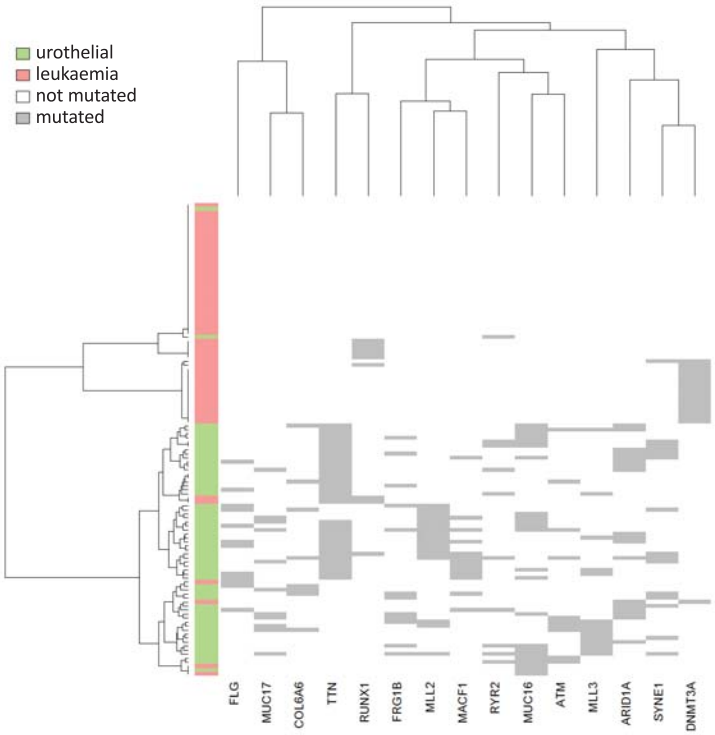
The rows represent samples and columns represent genes in this heatmap. The coloured column to the left of the heatmap indicates the class to which each sample belongs. The grey markers indicate in which gene each sample carried a functional mutation.The green column indicates sample age ranging from grey to dark green with increasing age.

Figure 4.1.6.9: Squamous cell carcinoma / leukemia heat map



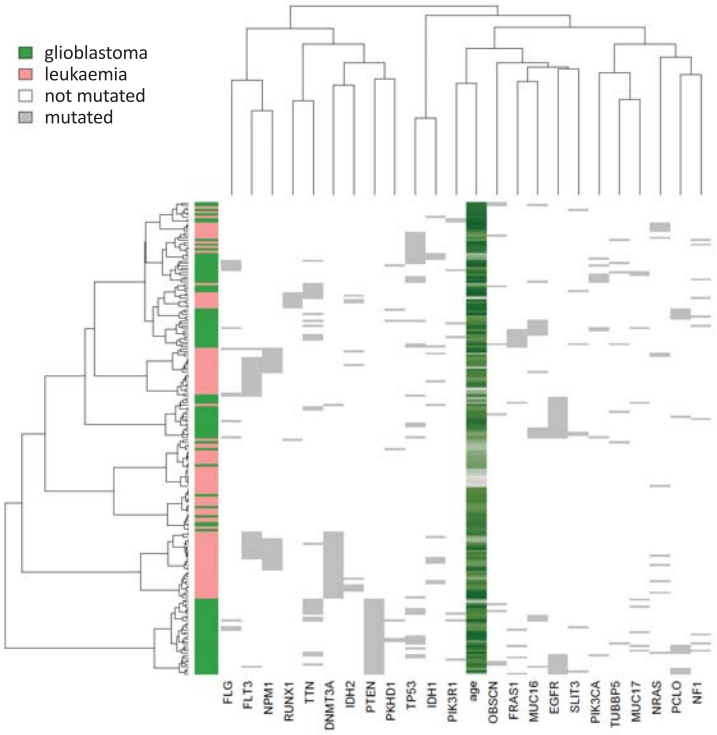
The rows represent samples and columns represent genes in this heatmap. The coloured column to the left of the heatmap indicates the class to which each sample belongs. The grey markers indicate in which gene each sample carried a functional mutation.

Figure 4.1.6.10: Urothelial / leukemia heat map



The rows represent samples and columns represent genes in this heatmap. The coloured column to the left of the heatmap indicates the class to which each sample belongs. The grey markers indicate in which gene each sample carried a functional mutation.

Figure 4.1.6.11: Glioblastoma / leukemia heat map



The rows represent samples and columns represent genes in this heatmap. The coloured column to the left of the heatmap indicates the class to which each sample belongs. The grey markers indicate in which gene each sample carried a functional mutation. The green column indicates sample age ranging from grey to dark green with increasing age.

Chapter 4.2

Working towards predicting the origin of cancers of unknown primary using whole exome sequence data.

4.2.1 Introduction

Cancers of unknown primary origin (CUP) make up around five percent of all diagnosed cancers and are the fourth leading cause of death from cancer (Pavlidis & Pentheroudakis, 2012). These cancers are typically aggressive metastases and patients can present with tumours at multiple locations with complex histology, making the diagnosis of the primary cancer type non-trivial. However, post mortem examinations can usually identify a primary tumour, suggesting that the current clinical practice is failing (Pentheroudakis et al., 2007). Recent work shows gene expression signatures of tumours of unknown primary origin can supply a likely tissue of origin classification (Tothill et al., 2005; Van Laar et al., 2009; Ramaswamy et al., 2001; Su et al., 2001).

Recent work on the Pan-Cancer dataset showed that cancers clustered according to their tissue of origin (Hoadley et al., 2014; Weinstein et al., 2013). The Pan-Cancer analysis also identified mutations which were almost exclusive to certain cancers; such as *APC* in colorectal cancer, and *VHL* mutations in kidney renal clear cell carcinoma (Kandoth et al., 2013a); a result which was confirmed in Chapter 4.1. A current theory is that the defining characteristics of CUP may be inherited from the tissue of primary origin (Dennis & Oien, 2005; Rosenfeld et al., 2008). Therefore, CUP that harbour *APC* mutations may be most likely originate from a colorectal adenocarcinoma. Whereas CUP that carry *VHL* mutations may be most likely to originate from a kidney renal clear cell carcinoma. Treatments known to be successful for certain cancers may also be successful for treating their metastases.

When diagnosing the origin of CUP clinicians should have available to them as much pathological, histological, and other molecular data as possible. In addition to gene expression classification tools for CUP (Tothill et al., 2005; Van Laar et al., 2009; Su et al., 2001), tools which could provide a putative origin for CUP, based on exome or genome sequencing, would be a useful aid to treatment choice. Tothill et al. (2013) used an exome sequence gene panel to classify CUP in a small dataset. Whole-exome sequence data has not yet been used to classify CUP.

Exome sequencing, and whole genome sequencing of CUP may also uncover mutations

for which there are known pharmacogenetics associations. Indeed there are now studies which have leveraged clinical exome sequencing of primary tumor biopsies to make decisions about which targeted treatments are most appropriate for a patient (Frampton et al., 2013). As sequencing costs continue to decrease, such sequence-based tests become economically viable because the high cost of a single sequencing run can be offset by the multitude of tests which can be completed using whole exome, or whole genome sequence data (Garraway & Lander, 2013).

In this thesis, I investigated the utility of a multi-class Random Forest classification tool to predict a primary site for CUP using whole exome sequence data comprising gene level features, variant type frequencies and transition and transversion variant frequencies. The tool was created and tested using a training and test set derived from the Pan-Cancer dataset. The Pan-Cancer dataset is not a comprehensive set of all cancers and so we cannot make accurate predictions for all cancers. As additional cancer exome sequence datasets become available through the cancer genome atlas we can use them to create CUP prediction tools that are more comprehensive. In the future this approach may also be adopted to predict the origin of circulating tumour cells, or cell-free DNA (Dawson et al., 2013) as a liquid biopsy tool.

4.2.2 Methods

I created an eleven class Random Forest classification model to classify cancers according to their tissue type. There is currently no suitable validation set in which to test the real world performance of this classifier. However, the Genomics England project plans to obtain the whole genome sequence of carcinomas of unknown primary origin. This will allow me to test how well the model agrees with the putative tissue of origin predicted by clinicians.

4.2.2.1 The Pan-Cancer dataset

I used the same Pan-Cancer dataset as in section 4.1.2.1 on page 105. I downloaded the Mutation Annotation Format (MAF) files from Sage Synapse on October 10th 2013. I removed mutations from the MAF file that were classified as 'silent'. Silent mutations are synonymous mutations, which do not lead to a change in the amino acid in the translated protein. I used the union of the lists of functionally mutated proteins across all samples to generate a matrix M of proteins P by samples S. Each element in $M[p, s]$, was set to 0 if there was no protein coding mutation for protein p, and sample s, or set to 1 if at least one protein coding mutation was present for protein p, and sample s.

I also computed the frequency of variant types and the six transition and transversion (TV/TS) SNV types. I extracted samples with available exome sequence and clinical data for age, gender and cancer stage. The Pan-Cancer dataset included Amyloid Myeloid Leukemia (AML) which is not suitable as an origin of a solid metastasis, and so AML samples were not included in the model building procedure.

4.2.2.2 Training set sampling

For each Pan-Cancer cancer type, two thirds of the samples were assigned to the training set for model building and one third to the test set. I then used two sampling strategies. The first strategy used a downsampling procedure to down-sample all classes to be of equal size to the smallest class in order to minimise the tendency of a classifier to defer to the

majority class (Barandela & Sã, 2003). The second strategy used both downsampling and the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002), and is outlined below.

4.2.2.2.1 SMOTE up-sampling

A combination of down-sampling, and the SMOTE (Chawla et al., 2002) up-sampling procedure was used to set the training set size of each of the ten cancer classes to be a multiple of the minimum training class size.

The two thirds training set size for each cancer class m was computed as the set of training set sizes M. The minimum training set size M_{min} was recorded. The up-sampling parameter u indicated the proportional increase in minority class training set size to be achieved through SMOTE over-sampling. Where $u = 2$, the minimum class size was to be doubled, and the balanced training set size to be achieved for each cancer class S was derived by $S = M_{min} \times u$. The set of cancer classes where the two thirds sample size was larger than the balanced training set size G was defined by $G = \{m \mid m \in M \text{ and } m > S\}$. For the G classes, a balanced training set size of S was achieved through random selection of S samples. The set of cancer classes with a two third sample size smaller than the balanced training set size L was defined by $L = \{m \mid m \in M \text{ and } m < S\}$. For each class $l \in L$ the number of samples to be generated by the SMOTE (Chawla et al., 2002) procedure $n \in N$ was found by subtracting the two thirds training set size for class l , M_l, from the balanced training set size S ($n = S - M_l$). For each class $l \in L$, the SMOTE procedure was used to create n additional samples, so that the training set size for all classes were equal to S.

The SMOTE Chawla et al. (2002) procedure works by generating new synthetic training set samples from a training set by using features drawn from a collection of P training samples. The similarity between the training set samples is represented in 'sample space' where the distance between the samples is inversely related to their similarity. The SMOTE procedure creates a new sample by picking two training samples, and uses combinations of the features of the two training samples to create a new, synthetic, training sample.

Table 4.2.2.1: Two third training set sizes for each cancer type used to build the CUP classifiers

	BLCA	BRCA	COADREAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	UCEC
N samples	61	511	149	183	204	278	105	116	210	164

The two thirds training test class sizes of each cancer type were imbalanced ($\chi^2(9, N = 1981) = 719.20,\ p = 2.2 \times 10^{-16}$).
BLCA = bladder, **BRCA** = breast, **COAD** = colon adenocarcinoma, **GBM**, = glioblastoma, **HNSC** = head and neck squamous cell carcinoma, **KIRC** = kidney, **LUAD** = lung adenocarcinoma, **LUSC** = lung squamous cell carcinoma, **OV** = ovarian cystadenocarcinoma, **READ** = rectum adenocarcinoma, **UCEC** = endometrial carcinoma.

This process results in a training set where classes either have been under-sampled without replacement, or up-sampled without replacement, in order to create a set of size S.

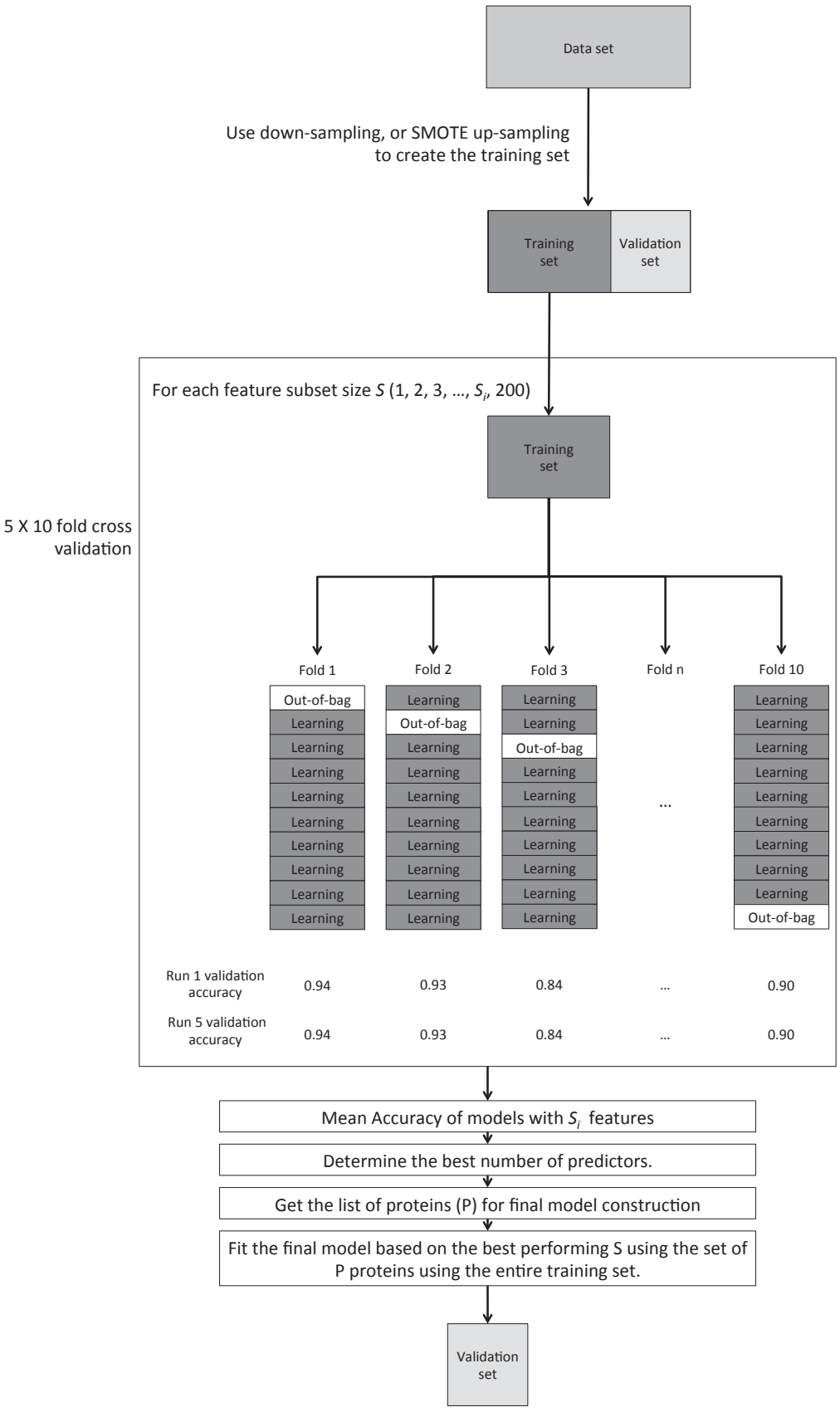
4.2.2.2.2 Mutated protein feature selection

Protein features were retained if they carried at least one protein coding mutation in five percent or more of any cancer class in the training set. By doing this I aimed to remove genes likely to contain passenger mutations from the protein features, and only use those features that were more likely to contain information that reflected biological characteristics of the tumour (Kandoth et al., 2013a).

4.2.2.3 Random Forest model building

The model building pipeline is shown in Figure 4.2.2.1. All Random Forest models were built using the *Caret* (Kuhn, 2008) package wrapper for *randomForest* in *R*. Models were built using recursive feature elimination with 5 x 10 fold cross validation in order to select the features for models composed of feature set sizes from 1-200 using feature set sizes of (1-50, 60, 70, 80, 90, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200). Mean classification accuracy was used to select the feature subset size, N, that resulted in the best performing models to use in final model building. The final model was built using the entire training set, and the top N ranked protein features. In an attempt to reduce the effect of over-fitting the classification models I chose to use the model with the smallest feature set size that performed within one percent of the best performing model.

Figure 4.2.2.1: Cancer of unknown origin classifier building pipeline



4.2.2.4 Classifier performance measures

For each cancer class, a true positive in the test data was defined as a correct classification of the cancer class label and a true negative was defined as a classification of any other class label as any other class label except the positive class. For example when considering colorectal adenocarcinomas, a true positive would be a colorectal adenocarcinoma being classified as a colorectal adenocarcinoma, and a true negative could be a lung adenocarcinoma classified as any other cancer class except colorectal adenocarcinoma. I used the area under the ROC curve (AUC) statistic for each class to measure classifier performance.

The analysis scripts used in this chapter can be found at https://github.com/SutherlandRuss/RS_PhD_scripts.

4.2.3 Results

4.2.3.1 The down-sampled classifier

The classifier which balanced the classes to be of equal size to the smallest training set of the 10 cancer classes performed with a mean accuracy of 72.5 percent (SD= 1.8) in the test dataset (Table 4.2.3.1). Across all classes, the mean classification sensitivity was 73.7 percent (SD= 15.7), and the mean classification specificity was 96.9 percent (SD=2.0). The selected model consisted of 23 features including all six of the transition and transversion frequency features, and seven of the variant type features excluding 'In-frame insertions', and 'non-stop mutations'. There were nine proteins in the model: *TP53*, *APC*, *PTEN*, *VHL*, *KRAS*, *PIK3CA*, *CTNNB1*, *CDKN2A*, and *PBRM1* (Table 4.2.3.2) presented in order of descending mean classification accuracy.

Each cancer class was predicted with at least 71.3 percent accuracy (Table 4.2.3.1), the lowest being found for head and neck squamous cell carcinoma and the highest for lung squamous cell carcinoma at 95 percent. By inspecting the receiver operating characteristic curve I observed that at all probabilities for being assigned to each class, the classifier performed well. Each class had an area under the curve (AUC) of at least 90 percent (Figure 4.2.3.2a).

4.2.3.2 The up-sampled classifier

The performance of the up-sampled classifier was similar to the down-sampled classifier (Table 4.2.3.1, and Figure 4.2.3.2b). The Smote up-sampled model included 30 features and performed with an accuracy of 74.1 percent (SE= 1.99). All six transition and transversion features were included in the model and seven of the variant type features excluding 'In-frame insertions', and 'non-stop mutations' (Table 4.2.3.2). Age and gender features were included in the model along with 15 protein features; *TP53*, *APC*, *KRAS*, *VHL*, *PTEN*, *PIK3CA*, *EGFR*, *PIK3R1*, *PBRM1*, *ARID1A*, *CTNNB1*, *CSMD3*, *ZFHX4*, *MUC16* and *FRG1B.*

There was no difference in classifier performance when AUC scores (Figure 4.2.3.1 and

Table 4.2.3.1: Test set performance statistics for the down-sampled and up-sampled CUP classifiers

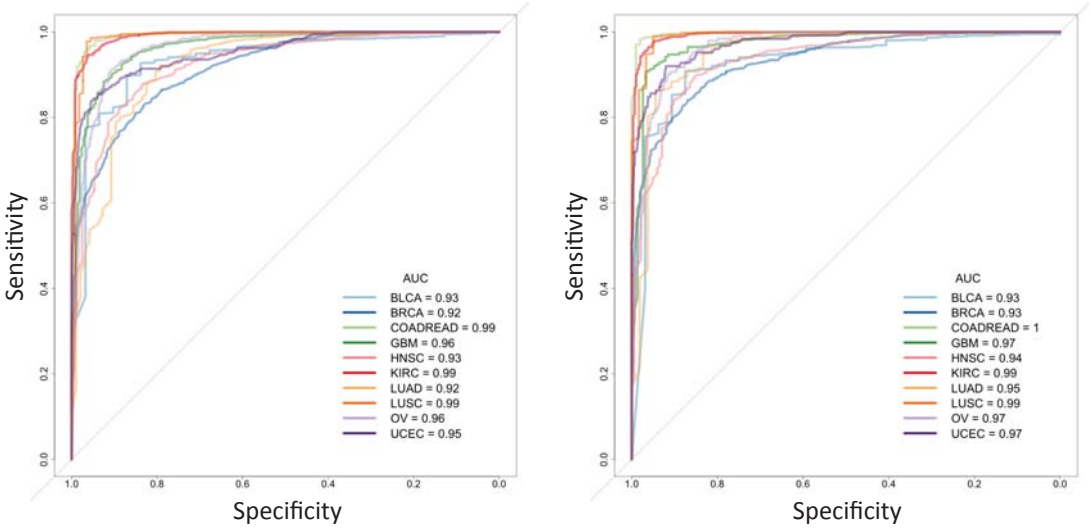
	Down-sampled classifier									
	BLCA	BRCA	COADREAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	UCEC
Sensitivity	0.61	0.61	0.90	0.79	0.45	0.91	0.58	0.91	0.87	0.75
Specificity	0.96	0.95	0.99	0.96	0.98	0.98	0.99	0.99	0.92	0.98
Positive Predictive Value	0.15	0.84	0.87	0.67	0.69	0.89	0.62	0.79	0.58	0.75
Negative Predictive Value	1.00	0.85	0.99	0.98	0.94	0.98	0.98	1.00	0.98	0.98
Prevalence	0.01	0.30	0.07	0.09	0.10	0.15	0.04	0.05	0.11	0.08
Detection Rate	0.01	0.18	0.06	0.07	0.05	0.14	0.02	0.04	0.09	0.06
Detection Prevalence	0.05	0.22	0.07	0.11	0.07	0.15	0.04	0.06	0.16	0.08
Accuracy	0.78	0.78	0.95	0.88	0.71	0.94	0.78	0.95	0.89	0.86
AUC	0.93	0.92	0.99	0.96	0.93	0.99	0.92	0.99	0.96	0.95

	Smote up-sampled classifier									
	BLCA	BRCA	COADREAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	UCEC
Sensitivity	0.50	0.63	0.91	0.87	0.48	0.92	0.63	0.93	0.89	0.75
Specificity	0.97	0.96	1.00	0.95	0.97	0.98	0.99	0.99	0.92	0.98
Positive Predictive Value	0.20	0.90	0.92	0.63	0.65	0.89	0.65	0.80	0.55	0.77
Negative Predictive Value	0.99	0.83	1.00	0.99	0.94	0.99	0.99	1.00	0.99	0.98
Prevalence	0.02	0.35	0.06	0.08	0.10	0.16	0.030	0.03	0.11	0.07
Detection Rate	0.01	0.22	0.05	0.07	0.05	0.15	0.02	0.03	0.09	0.05
Detection Prevalence	0.04	0.25	0.06	0.11	0.07	0.17	0.03	0.04	0.17	0.07
Accuracy	0.73	0.80	0.95	0.91	0.73	0.95	0.81	0.96	0.90	0.87
AUC	0.93	0.93	1.00	0.97	0.94	1.00	0.95	1.00	0.97	0.97

The classifier that used down-sampling to balance classes in the training set (**Down-sampled classifier**) performed comparably to the classifier that used a combination of down-sampling and smote up-sampling (**smote up-sampled classifier**).

BLCA = bladder, **BRCA** = breast, **COAD** = colon adenocarcinoma, **GBM**, = glioblastoma, **HNSC** = head and neck squamous cell carcinoma, **KIRC** = kidney, **LUAD** = lung adenocarcinoma, **LUSC** = lung squamous cell carcinoma, **OV** = ovarian cystadenocarcinoma, **READ** = rectum adenocarcinoma, **UCEC** = endometrial carcinoma.

Figure 4.2.3.1: Cancer of unknown origin classification ROC curves



(a) Down-sampled training set (b) Smote up-sampled training set

The up-sampled classification model showed equal, or marginally larger AUC values in comparison to the down-sampled model.

Table 4.2.3.1) for each cancer class were compared between the down-sampled models and the smote up-sampled models using a bootstrap permutation method (Hanley & McNeil, 1983) and adjusted for multiple testing (BLCA [$D = 0.20, p = 0.84$], BRCA= [$D = -2.44, p = 0.15$], COADREAD= [$D = -1.22, p = 0.22$], GBM= [$D = -0.77, p = 0.44$], HNSC= [$D = -1.11, p = 0.26$], KIRC= [$D = -0.87, p = 0.38$], LUAD= [$D = -1.52, p = 0.13$], LUSC= [$D = -0.45, p = 0.65$], OV= [$D = -0.57, p = 0.56$], and UCEC= [$D = -1.99, p = 0.05$]).

Table 4.2.3.2: Random Forest classifier feature importance

Down-sampled classifier								
	Features	**Importance**		**Features**	**Importance**		**Features**	**Importance**
1	C>T_G>A	43.73	**11**	Missense_Mutation	22.39	**21**	CDKN2A	13.67
2	TP53	36.59	**12**	A>T_T>A	21.34	**22**	A>C_T>G	11.90
3	APC	34.10	**13**	Frame_Shift_Ins	21.26	**23**	PBRM1	8.21
4	C>A_G>T	33.36	**14**	Frame_Shift_Del	21.03			
5	gender	31.29	**15**	KRAS	20.93			
6	RNA	29.13	**16**	PIK3CA	19.44			
7	C>G_G>C	27.92	**17**	Nonsense_Mutation	18.93			
8	PTEN	24.74	**18**	CTNNB1	18.60			
9	A>G_T>C	24.03	**19**	Splice_Site	17.79			
10	VHL	23.89	**20**	In_Frame_Del	16.58			

SMOTE up-sampled classifier								
	Features	**Importance**		**Features**	**Importance**		**Features**	**Importance**
1	TP53	49.80	**11**	PTEN	28.92	**21**	CTNNB1	19.07
2	C>T_G>A	47.27	**12**	A>T_T>A	28.89	**22**	EGFR	17.42
3	gender	47.00	**13**	PIK3CA	26.88	**23**	age	16.29
4	C>A_G>T	41.90	**14**	In_Frame_Del	26.53	**24**	ARID1A	14.55
5	APC	41.14	**15**	Missense_Mutation	26.17	**25**	CSMD3	13.51
6	RNA	39.70	**16**	VHL	25.83	**26**	FRG1B	13.41
7	C>G_G>C	37.75	**17**	A>C_T>G	24.14	**27**	PBRM1	12.75
8	A>G_T>C	34.38	**18**	KRAS	23.01	**28**	PIK3R1	12.66
9	Frame_Shift_Ins	31.18	**19**	Nonsense_Mutation	20.50	**29**	ZFHX4	9.69
10	Frame_Shift_Del	29.12	**20**	Splice_Site	19.97	**30**	MUC16	9.64

The variables included in each Random Forest model are ranked according to their mean classification accuracy in the training set.

4.2.4 Discussion

The classification models contained proteins which are exclusively mutated in certain Pan-Cancer cancer types. *PBRM1* and *VHL* mutations are specific to renal cell carcinoma (Kandoth et al., 2013a). *APC* mutations are mostly specific to colorectal cancer in this dataset (Kandoth et al., 2013a). *CDNK2A* is most commonly mutated in the squamous cell carcinomas; head and neck squamous cell carcinoma, and lung squamous cell carcinoma (Kandoth et al., 2013a). Mutations in *PTEN*, and *PIK3CA* are most common in endometrial carcinoma. *KRAS* mutations are most common in colorectal cancer and lung adenocarcinoma.

The CUP Random Forest model built on down-sampled data performed equally to the classifiers built using the SMOTE up-sampling procedure. The features included in the model built using the SMOTE up-sampled data included two genes; *CSMD3*, and *MUC16*, which have been found to be significantly mutated in some TCGA studies, but are thought to be false positives (Lawrence et al., 2013). The down-sampled model is preferred because the

protein features are more biologically plausible than those included in the up-sampled model.

The Random Forest method is limited by the number of cancer classes included in the training set. I cannot make predictions about the origin of a cancer, if the tissue of its true origin was not included in training set. In addition, a sample is always assigned to the class with the highest classification probability. This may increase the false positive rate and reduce classification specificity. An alternative would be to include an *unknown class* category. By setting a threshold for the probability at which any sample can be assigned to a class, the specificity of the classifier may increase, because samples which have a low probability of belonging to any of the 10 cancer type categories would be assigned to the *unknown* category. In addition, the R implementation of Random Forest is limited to 32 categories, allowing for only 32 outcome classes. It may be preferable to use a k-nearest neighbour, or k-medoids approach to build a classifier which will have no limitation on the number of outcome classes to which an unknown sample may be assigned.

4.2.5 Conclusion

By providing clinicians with a predicted tissue of origin based upon exome sequence mutations, patient treatment decisions could be aided. The numbers of CUP that have been exome sequenced is too small at present to draw strong conclusions about this approach. However, within the next two years Genomics England plan to whole genome sequence many stage four CUP. This will enable researchers to develop tools to predict the primary origin of late stage metastatic cancers based on NGS data and provide some guidance to clinicians regarding patient care.

There is further utility of this approach for tumour cells that circulate in blood, along with cell free tumour DNA. A machine learning classification approach, such as I have taken here, may be able to assign a tissue of primary origin where circulating tumour cells are detected from a blood test, without the need for MRI scan time to establish the primary site of the cancer.

Chapter 5

Candidate disease gene prioritisation for complex diseases using network-based methods

Univariate tests to identify significantly mutated genes in cancer, or to identify variants associated with complex traits have been successful in identifying some of the genetic causes of cancer and complex diseases. However, not all of the genes that explain cancer or the heritable component of other complex diseases have been discovered. One reason for this may be that these diseases are heterogeneous in terms of disease subtypes, where each subtype has distinct genetic causes. Another reason may be that complex diseases are genetically heterogeneous, with causal variants distributed among genes that operate as part of a pathway. Univariate tests, may be underpowered to detect these associations (Leiserson et al., 2013b).

Network-based tests may be able to detect disease associations at the pathway level and suggest additional candidate disease genes. This chapter presents two investigations in to integrating genomic data with protein interaction network data to provide additional candidate disease genes. In Chapter 5.1 I developed the 'k-pseudo cliques analysis' and analysed the PanCancer colorectal cancer exome sequence data (Kandoth et al., 2013a). In Chapter 5.2 I used the Region Growing Analysis method developed by Lehne (2011) and Christopher Tebbe to analyse rheumatoid arthritis single nucleotide polymorphism data from the Wellcome Trust Case-Control Consortium genome-wide association study (Burton et al., 2007).

Chapter 5.1

Network-based disease gene prioritisation using colorectal cancer exome sequence data

5.1.1 Introduction

Cancer is a complex and heterogeneous disease in which DNA mutation is the initial cause of cancer and it is also part of the cancer phenotype. The mutations that initiate cancer progression are known as *driver* mutations, and occur at a high frequency in exome sequence tumour data (Ding et al., 2014). Genes carrying driver mutations, sometimes called *cancer genes*, are mutated more frequently across samples (Kandoth et al., 2013a) and many algorithms exploit this property to measure the degree to which genes are significantly mutated (Lawrence et al., 2013; Dees et al., 2012). In some cancers, such as colorectal cancer, significantly mutated gene products also physically interact with one another (Leiserson et al., 2013a; Ciriello et al., 2012; Miller et al., 2011). This suggests that, at least for some cancers, multiple genes can contribute to the cancer, and they can be part of the same functional pathway.

Cancer, as well as other complex diseases, may arise due to the perturbation of pathways. A perturbation to an important pathway, through knockout or gain of function mutation, of a subset of any pathway members may cause the same disease phenotype. The example of Acne Inversa shows this effect in a monogenic disease where a mutation in any one of three proteins adjacent in a protein protein interaction (PPI) network were found to cause the disease (Wang et al., 2010). Genes that contribute to the same, or similar diseases, are also closer than expected in PPI networks (Barrenas et al., 2009; Gandhi et al., 2006; Goh et al., 2007).

In single gene analysis there is the possibility that a pathway that contributes to cancer through the mutation of any one of the pathway members may not be discovered due to sample heterogeneity (Vogelstein et al., 2013). The multiple hypothesis testing burden at the pathway level should be lower than at the protein level because, by definition, there are fewer pathways than there are genes. Therefore, the power to identify plausible disease gene candidates should be higher when taking a pathway approach, in comparison to a single gene approach.

The discovery of pathways involved in cancer can be approached from a graph modelling

perspective. For example, the identification of pathways contributing to cancer can be represented as a community detection problem in a PPI network. There is no universally accepted definition of a community in the network literature. In PPI networks one way of defining the community detection problem is to ask "which groups of proteins are part of the same functional pathway?". Of course, this definition is loose, but it is applicable to most of the approaches that have been taken to discover communities in protein interaction networks (Leiserson et al., 2013b). The community detection problem itself can be approached in numerous ways. Some approaches take a seed and disperse approach, by *growing* pathways based on *seed* vertices expanding the pathway based on maximising a statistic (Chuang et al., 2007; Leung et al., 2014; Ozgun et al., 2014). Approaches such as Markov Clustering (MCL) (Enright et al., 2002), and the Girvan-Newman algorithm (Girvan & Newman, 2002) partition the graph in to discrete subgraphs. Whereas, other approaches such as: clique percolation (Palla et al., 2005, 2007), and pseudo clique enumeration (Georgii et al., 2009; Tsuda & Georgii, 2009) identify overlapping subgraphs. In this study I identified overlapping communities of vertices using pseudo clique enumeration (Georgii et al., 2009), and a k-pseudo clique search algorithm (Palla et al., 2005) to find all k-pseudo cliques in the Human Protein Protein Interaction network 2 (HuPPI2) (Lehne & Schlitt, 2009). I then investigated whether the resultant communities were enriched with genes that had lower than expected p-values for a gene-level test.

Two main tests have been used to identify significantly mutated genes in TCGA cancer datasets; MuSiC (Dees et al., 2012), and the set of MutSig algorithms (Lawrence et al., 2013; Banerji et al., 2012). MutsigCV (Lawrence et al., 2013) is a method that is used to estimate the extent to which each gene is significantly mutated in a set of tumour samples based upon the somatic mutations, which are unique to the tumour. The earliest iterations of the mutsig algorithm measured whether each gene was significantly mutated by calculating a background mutation rate which was assumed to be constant across the genome. However, the background mutation rate is correlated with gene expression levels (Pleasance et al., 2010), and DNA replication time (Stamatoyannopoulos et al., 2009). MutsigCV extends

the mutsig method by estimating the background mutation rate for each gene based upon the number of silent and mutations within each gene. If the numbers of mutations in the genes are not sufficient to properly estimate the background mutation rate, the k-nearest neighbours of the gene are used to calculate the background mutation rate. This is adjusted for the mean expression value of the gene across all tissues in the cancer cell line encyclopedia, and the DNA replication time. Genes that are replicated late during mitosis have a higher mutation rate than those that are replicated at the start of the DNA replication cycle. I used the MutSigCV (Lawrence et al., 2013) algorithm to assign a p-value to each gene that indicated the degree to which each gene was significantly mutated. I then used the PPI network, as prior biological information, and the Dense Module Enumeration (DME) implementation of the pseudo clique enumeration algorithm (Georgii et al., 2009; Tsuda & Georgii, 2009) to identify gene sets, or communities, that were concentrated with low p-values.

The significant gene sets, corresponding to the communities, represented functionally related genes that were not significantly mutated at the gene level, but were part of a significantly mutated subnetwork. This approach identified gene sets that contained genes that were known to contribute to colorectal cancer and additional candidate genes carrying functional mutations that may be good candidates for wet-lab cancer study and drug repurposing. The k-pseudo cliques analysis method is not limited to analysis of cancer sequencing data. It has many applications in the field of complex disease genetics and can be used to analyse any genome-wide gene-level test statistic data.

5.1.2 Methods

5.1.2.1 Mutation and network data

I downloaded colorectal cancer mutation annotation format (MAF) files and clinical data from Sage Synapse (accession syn1710680) on October 3rd 2013. These files contained information in a VCF-like format. Each record (line) corresponded to a mutation. The sample ID, chromosome, base-position start, base-position stop, variant classification, and variant type data correspond to columns and were available for all mutations. I retained 219 samples for which age, gender, and mutation data were available. I removed 24 hypermutated samples which contained more than 500 somatic mutations (Kandoth et al., 2013a). I ignored silent, synonymous, mutations, but retained protein coding mutations. For each sample I labelled the genes that carried at least one functional mutation to create the binary mutation matrix as defined in earlier chapters.

I used the HuPPi2 network (Lehne & Schlitt, 2009) as prior biological information and to define the k-pseudo cliques.

5.1.2.2 MutSigCV analysis

I used MutSigCV to identify the genes that were significantly mutated in colorectal cancer adjusted for gene length, expression level, and DNA replication time (Lawrence et al., 2013). I ran the analysis using the GenePattern Broad Institute server (Reich et al., 2006) using the default parameters.

5.1.2.3 The HuPPI2 network

The HuPPI2 network (Lehne & Schlitt, 2009) is an example of a meta-protein interaction network. It was composed of the union of the protein interactions present in six protein interaction databases; BioGRID, MINT, DIP, IntAct, and HPRD. The network was created in May 2008. Only protein interactions classed as *physical* were used. Furthermore, interactions were retained only if they had been independently corroborated by two publications.

5.1.2.4 Enumeration of pseudo cliques

Pseudo cliques are a type of graph community defined by their density of edges among member vertices. A connected subgraph was defined as a pseudo clique if the ratio of the number of edges to the maximal number of edges, density, was above a user defined density threshold, α[0,1]. I used five increasingly relaxed α thresholds to define pseudo cliques: 0.95, 0.90, 0.85, 0.80, and 0.75. The number of vertices belonging to each pseudo clique denoted the pseudo clique size. I used Tsudo and Georgii's DME program (Tsuda & Georgii, 2009; Georgii et al., 2009) to enumerate all of the pseudo cliques within a weighted graph using a reverse search. I then loaded each DME results file in to R along with the mutation data and the PPI network for further analysis using igraph (Csárdi & Nepusz, 2006).

5.1.2.5 K-pseudo clique definition

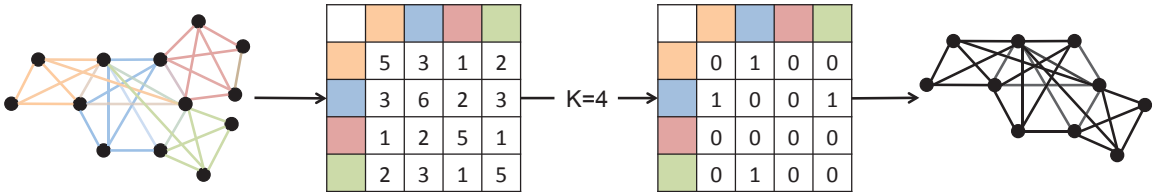
The DME algorithm identifies pseudo cliques of maximal size that satisfy the density threshold. The result of this is that although many pseudo cliques may be discovered, their gene sets may only differ by a small number of genes. Therefore, I reduced the number of gene sets and reduced the correlation between the gene sets by generalising the clique percolation method (Palla et al., 2005) to pseudo cliques in order to identify k-pseudo clique communities.

In this algorithm the parameter 'k' defined the minimum size of a pseudo clique that could be included in a k-pseudo clique community. For the set of pseudo cliques discovered at each density threshold and for each value of k, I created the pseudo clique overlap matrix. This matrix was analogous to the clique overlap matrix from the CFinder algorithm (Palla et al., 2005). Element E_{ij} indicated the number of vertices, or proteins, shared by pseudo cliques i and j. For each value of k I set the diagonal elements of the pseudo clique overlap matrix to 0 and subtracted $k-1$ from all off diagonal elements. This modified pseudo clique overlap matrix represented an adjacency matrix of the pseudo cliques of at least size k that intersected by at least $k-1$ proteins. I then used a component decomposition of the adjacency matrix to identify the k-pseudo cliques. For each k-pseudo clique, its gene set was found from the intersection of its pseudo clique members. Each k-pseudo clique gene set

was then taken forward for further analysis. Figure 5.1.2.1 depicts the process of identifying k-pseudo cliques.

My contribution was to apply the clique percolation method (Palla et al., 2005) to pseudo cliques, to identify highly interacting communities and to see whether those communities were enriched with genes with lower than expected p-values for a gene-level test.

Figure 5.1.2.1: The k-pseudo clique search process
Three pseudo cliques that overlap by 3 vertices are combined in to a single k-pseudo clique.



There are four pseudo cliques where a path can be traced between each pseudo clique to all other pseudo cliques by intervening pseudo cliques. The vertices that belong to each pseudo clique are indicated by the colour of the edges. The number of vertices shared by the pseudo cliques is represented as the pseudo clique - pseudo clique overlap matrix. The orange pseudo clique and the blue pseudo clique share three vertices. In order to group the pseudo cliques I modify the pseudo clique - pseudo clique overlap matrix. I define k as the minimum size of pseudo clique I consider. I set all diagonoal elements in the matrix to 0 and subtract $k-1$ from all off diagonal elements. The matrix now indicates which of the pseudo cliques of size 4 overlap by at least 3 vertices. This is an adjacency matrix representation of a network of pseudo cliques. I perform a component decomposition on this matrix to extract the list of all k-peudo cliques. This process is conducted in order to reduce the number of gene set tests I measure for their enrichment with low gene level test p-values.

5.1.2.6 Network permutation

I mapped MutSigCV scores for each gene to corresponding vertices on HuPPI2 (Lehne & Schlitt, 2009) network. For each pseudo clique I recorded the median MutSigCV p-value of the belonging genes. In order to establish whether the k-pseudo cliques I discovered were enriched for genes contributing to colorectal cancer I used a permutation test. I randomly reassigned the vertex labels in the network, leaving the k-pseudo clique structures unchanged, and recomputed the median MutSigCV p-value of each k-pseudo clique 10 000 times. A k-pseudo clique was significantly enriched with disease contributing genes if its median MutSIgCV p-value from the non-permuted network was ranked in the top five percent (500)

of its median MutSigCV p-values from the 10 000 network permutations.

The permutation test was required to establish whether any enrichment in gene-level MutsigCV results I saw in the k-pseudo cliques was due to network structure (a chance relationship), or because of the functional relationships between proteins encoded by edges in the network. Any network-based permutation test using a PPI network must break the functional relationships between proteins in the PPI network. By shuffling the vertex labels, the protein names, the functional relationships indicated by the edges between proteins in the PPI network would be broken. I used a degree constrained label reassignment procedure to shuffle the vertex labels in the PPI network. It was possible to reassign vertex labels 'within degree', but to do so meant vertices of unique degree would never have been reassigned to other vertices. In order to ensure that vertex labels were reassigned I tried two approaches based on reassigning vertex labels constrained by vertex degree.

5.1.2.6.1 Absolute degree difference constrained label reassignment

In this algorithm all vertices of the same degree were simultaneously assigned a new label. First of all, a random degree, d, was selected from the set of vertex degrees D. I ranked the set of all vertices, X, according to their absolute degree difference $|\triangle_d|$ to d. For each randomly selected d the set of vertices of degree d was defined as d_n. The new vertex labels $x_n \subseteq X$ for d_n were selected from the set of all available vertices X according to the $|\triangle_d|$ to d using a single-tailed normal distribution with a mean $|\triangle_d|$ of 0 and standard deviation of 5. This ensured that all vertices may be selected, and that the labels of vertices with a degree of d were reassigned with vertex labels of vertices with a degree of d+-10 with a probability of 0.95. The selected vertex names x_n were removed from the sampling pool of vertices X. A new d was then randomly selected and the above steps were repeated until all values of X had been re-assigned.

Although this method of degree-constrained network permutation did ensure that vertex names were likely to be reassigned with the names of vertices of similar degree, there were problems. The label reassignment procedure aimed to reassign the names of most vertices

to vertices of the same degree. Under this sampling procedure vertices of high degree, which were rare, were less likely to be re-assigned another vertex name. In addition, vertices of degree d were more likely to be assigned the names of vertices with an absolute degree difference larger than zero.

5.1.2.6.2 Degree constrained label reassignment

This algorithm randomly re-assigned vertex names of a graph constrained by the degree of the vertex to be reassigned. It is heavily based upon the degree constrained label shuffling algorithm of Lehne 2011, where two vertices are selected and their labels are swapped. My method selects two vertices A and B, and assigns the label of vertex B to vertex A. The label of vertex A is retained in the pool of vertex names to be re-assigned. First of all, a random vertex v was selected. All vertices were ranked and ordered according to their absolute degree difference $|\triangle_v|$ to v. I randomised the ranks of vertices of identical degree, so as to eliminate the alphabetical bias when using R's sorting procedure. New vertex labels were selected according to a sampling probability. The sampling probability for each vertex was based on the rank position of each vertex. Using a single-tailed normal distribution with a mean $|\triangle_v|$ of 0 and standard deviation of 5 I selected a new vertex $x \subseteq X$ and reassigned the vertex label and associated metadata to that of vertex v. This procedure ensured that all vertices were available for selection, and that the metadata of a vertex was likely to be reassigned to another vertex with 0.95 probability that the two vertices were less than 10 positions apart in the vector of vertices sorted by $|\triangle_v|$. I used the degree constrained label reassignment method 10 000 times to create 10 000 network permutations.

I decided to use vertex label reassignment as the network permutation method, rather than edge shuffling, or a combination of both methods, because of limitations of the DME program. The DME program may take weeks to run on a relatively small network such as the HuPPi2 at a low α threshold of 0.75. Degree constrained vertex label reassignment retained the topology of the network while controlling for vertex degree effects, and breaking any functional links between adjacent proteins in the network. The resulting permuted networks

and the original network were isomorphic. Pseudo cliques identified in the observed network were applicable to the permuted networks. Using an alternative edge shuffling permutation would have changed the topology of the observed network meaning that permuted networks and the observed network would not be isomorphic. If edge shuffling was used to permute the observed network the pseudo cliques would have changed, and the very community structures I wanted to test would have been broken apart. Degree constrained label re-assignment may not be the most comprehensive network permutation, but it provided an acceptable way to create permuted networks when it was important that the permuted networks and the original network be isomorphic. It is of course conceivable that any vertex metric other than vertex-degree may be used to constrain vertex label permutation. However, I have not addressed that in this thesis.

5.1.2.7 KEGG enrichment analysis

I aimed to understand if the significant results from the k-pseudo clique network-based approach were more enriched for cancer related genes in when compared with results from a univariate, single protein, approach. As such, I compared the KEGG pathway enrichment p-value for a gene set obtained from the k-pseudo clique network-based approach at FDR<0.1 (Benjamini & Hochberg, 1995) to the p-value obtained by the MutSigCV gene set significant at FDR<0.1.

At each α threshold and k I extracted gene lists corresponding to the union of the k-pseudo cliques that were significantly mutated at FDR<0.1. From the univariate MutSigCV test I extracted the list of genes that were significantly mutated at FDR<0.1. I performed KEGG Disease Pathway enrichment analysis on each of the resulting gene sets using the Webgestlat online tool (Wang et al., 2013) and the hypergeomteric test. I restricted tests to pathways containing four or more genes and restricted outputs to include only the gene sets that passed FDR<0.1. For each α threshold and k the colorectal cancer pathway, and pathways in cancer enrichment p-values were compared to the enrichment p-value for the univariate, MutSigCV, test gene set. The k-pseudo clique method was considered to add

additional information if the enrichment p-value of the gene sets derived from a k-pseudo clique analysis was smaller than that obtained from the MutSigCV gene set.

Each k-pseudo clique that contained at least ten proteins and was significantly mutated at FDR<0.1 was subjected to KEGG disease pathway enrichment as above. The disease pathway in which each k-pseudo clique was most enriched was recorded.

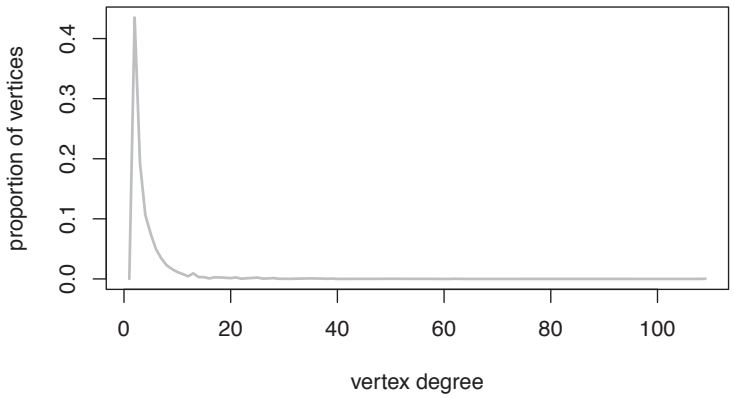
The analysis scripts used in this chapter can be found at https://github.com/SutherlandRuss/RS_PhD_scripts.

5.1.3 Results

5.1.3.1 Network statistics

The HuPPI2 network was composed of 3666 proteins (vertices), and contains 6187 pairwise interactions between the proteins (edges). It had a mean degree of 3.38 (sd=5.32). The average path length was 5.88, meaning that most vertices can reach any other vertex in the network by traversing 6 edges. The degree distribution was skewed towards 0, and was characteristic of other real world networks where most vertices are of low degree and very few vertices are of high degree (Figure 5.1.3.1).

Figure 5.1.3.1: HuPPI2 network degree distribution



Vertices of low degree are common, vertices of degree d become less frequent as d increases.

5.1.3.2 Network permutation results

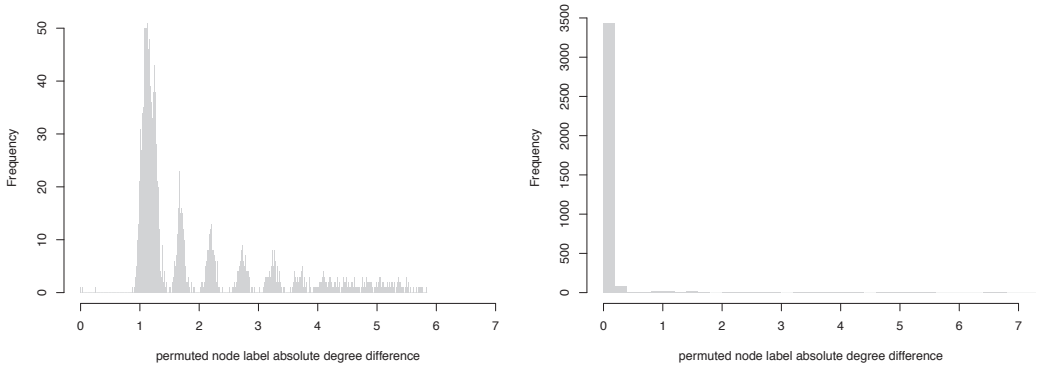
For 1000 vertex label re-assignment permutations I recorded the mean degree of the permuted vertex label in the original network. I tested the two label reassignment procedures using 1000 test permutations prior to deciding upon the final permutation procedure to use in this analysis. The absolute degree difference constrained label reassignment procedure did not

preferentially assign vertex names to vertices of the same degree, it was more likely for a vertex to be assigned the name of a vertex with an absolute degree difference of 1 or more (Figure 5.1.3.2a). The implementation of degree constrained label reassignment achieved the aim of preferentially randomly assigning vertex names to vertices of similar degree. For 1000 permutations the vast majority of vertex names were reassigned to vertices of the same degree (Figure 5.1.3.2b). All permutation testing was conducted using the degree constrained reassignment network permutation method.

Figure 5.1.3.2: Distributions of the absolute degree difference of permuted vertex labels and original vertex labels using two permutation methods.

(a) Absolute degree difference constrained label reassignment.

(b) Degree constrained label reassignment.

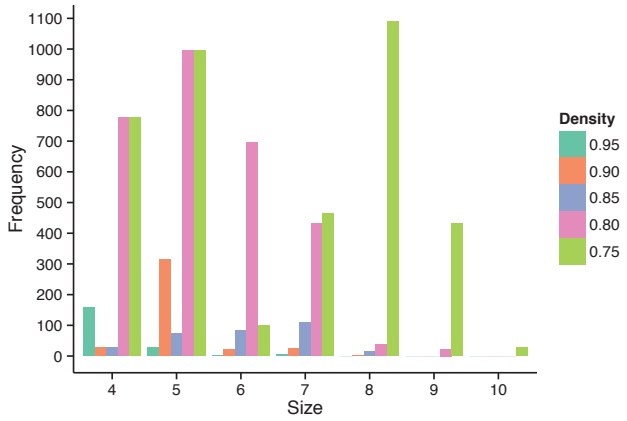


For each vertex the mean absolute degree difference between the original label and 1000 permuted labels was calculated and visualised as a histogram. When using absolute degree difference constrained label reassignment (Figure 5.1.3.2a) few vertices had a tendency to swap label with vertices of exactly the same degree, indicated by a mean absolute degree difference between 0 and 1. When using degree constrained label reassignment (Figure 5.1.3.2b) the majority of vertices had a tendency to swap vertex labels with vertices of the same degree, indicated by a mean absolute degree difference of 0.

5.1.3.3 Number and type of pseudo cliques identified at each density threshold

The frequency distribution of pseudo cliques at the density thresholds; 0.95, 0.90, 0.85, 0.80, and 0.75; are shown in Figure 5.1.3.3. As the density threshold decreased, the number of pseudo cliques identified increased, as did the size of the pseudo cliques. Within each density threshold the frequency distribution of k-pseudo clique size is shown in Figure 5.1.3.4.

Figure 5.1.3.3: Pseudo-clique size frequencies from change across density thresholds α=0.95 to α=0.75.



As the α parameter decreases, the size of discovered pseudo cliques increases. At α=0.80 there is around a ten-fold increase in the number of pseudo cliques discovered in comparison to α=0.85.

5.1.3.4 Median univariate test statistic permutation tests

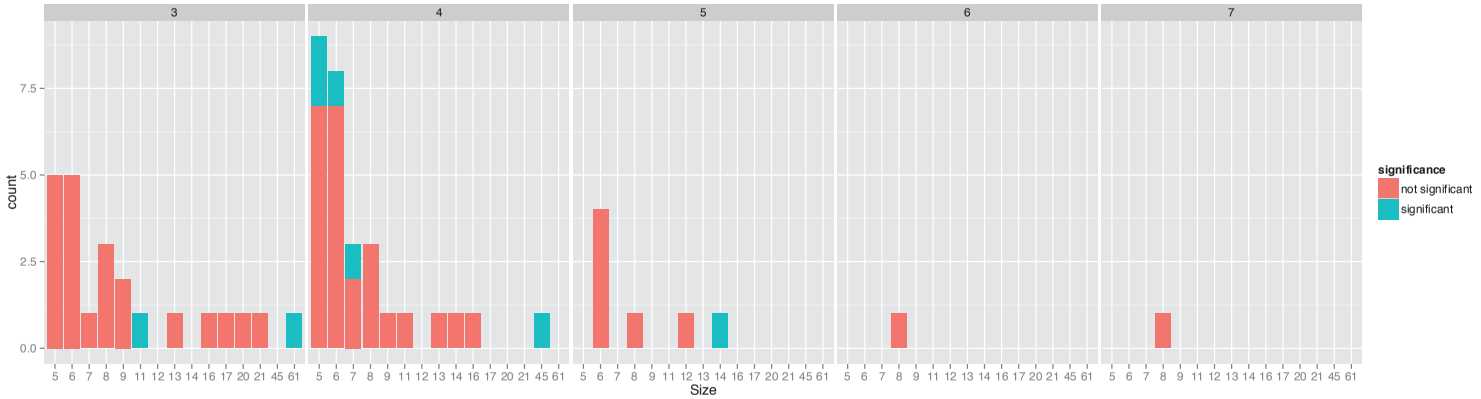
Across the 25 different combinations of *k* (3, 4, 5, 6, and 7) and α (0.95, 0.90, 0.85, 0.80, and 0.75) parameters I discovered 12 k-pseudo cliques that contained a lower than expected median MutSigCV pvalue at FDR>0.1. Results are shown in Table 5.1.3.1. The 12 pseudo cliques included 87 unique genes.

5.1.3.5 KEGG enrichment analysis

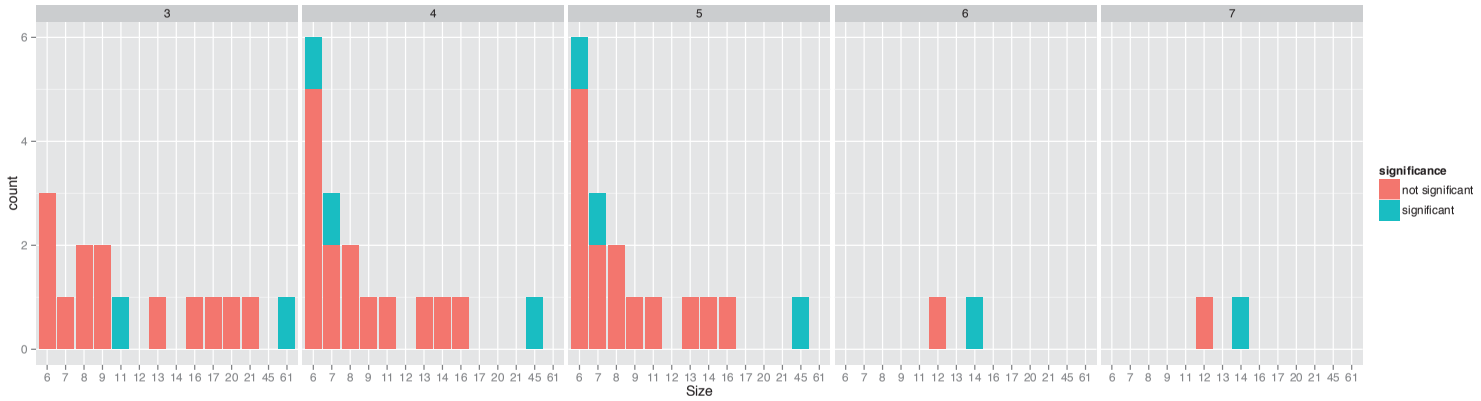
I wanted to understand whether the k-pseudo clique network-based analysis improved the detection of known cancer causing genes in comparison to a univariate test. None of the k-pseudo cliques analysis derived gene sets were more enriched in the colorectal cancer KEGG pathway than was the MutSigCV gene set at FDR<0.1 (Table 5.1.3.2).

In the uniprotein MutSigCV analysis, I found 12 genes *(KRAS*, *NRAS*, *TP53*, *APC*, *SMAD4*, *FBXW7*, *SMAD2*, *BRAF*, *TNFRSF10C*, *TCF7L2*, *C17orf97*, and *CTNNB1*) to be significantly mutated at an FDR of <0.1. Eight of these genes *(KRAS*, *TP53*, *APC*,

Figure 5.1.3.4: Histograms of significant k-pseudo cliques



(a) α=0.95

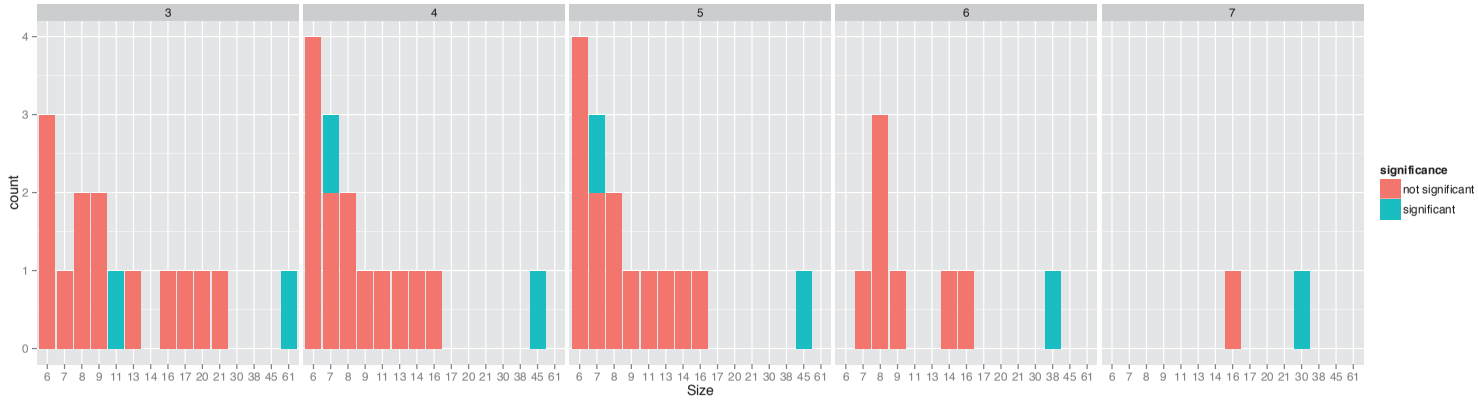


(b) α=0.90

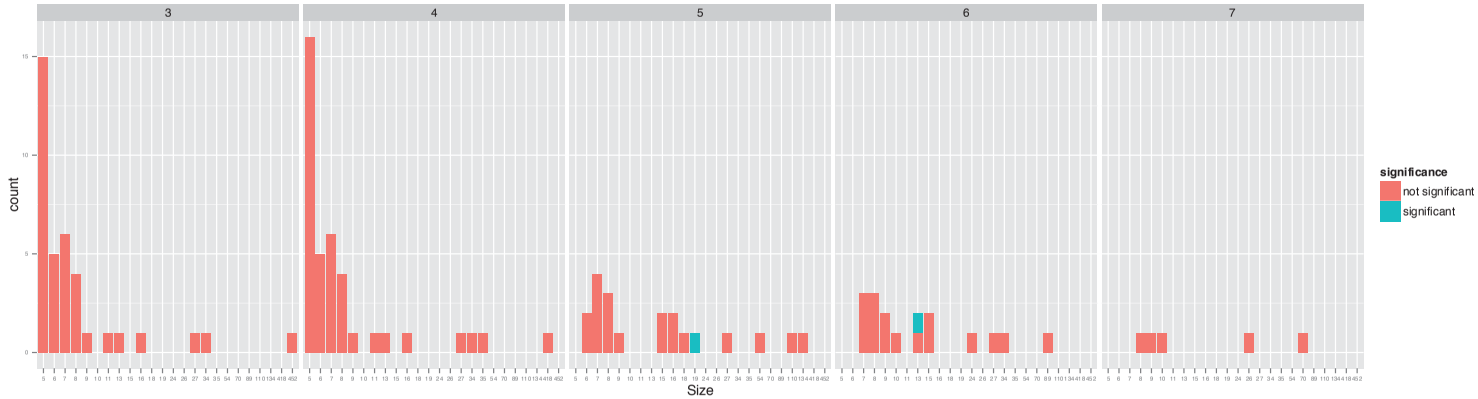
The x-axis represents the size of k-pseudo cliques found at each k setting. The significant pseudo cliques are indicated in green. In Figures 5.1.3.5a and 5.1.3.5b at each k setting smallest k-pseudo cliques were the most numerous and the number of k-pseudo cliques discovered decreased as k-pseudo clique size increased. In Figure 5.1.3.5b, the $k = 4$, and $k = 5$ parameter settings produced identical results.

Figure 5.1.3.4: k-pseudo clique size historgrams continued.

(c) α=0.85



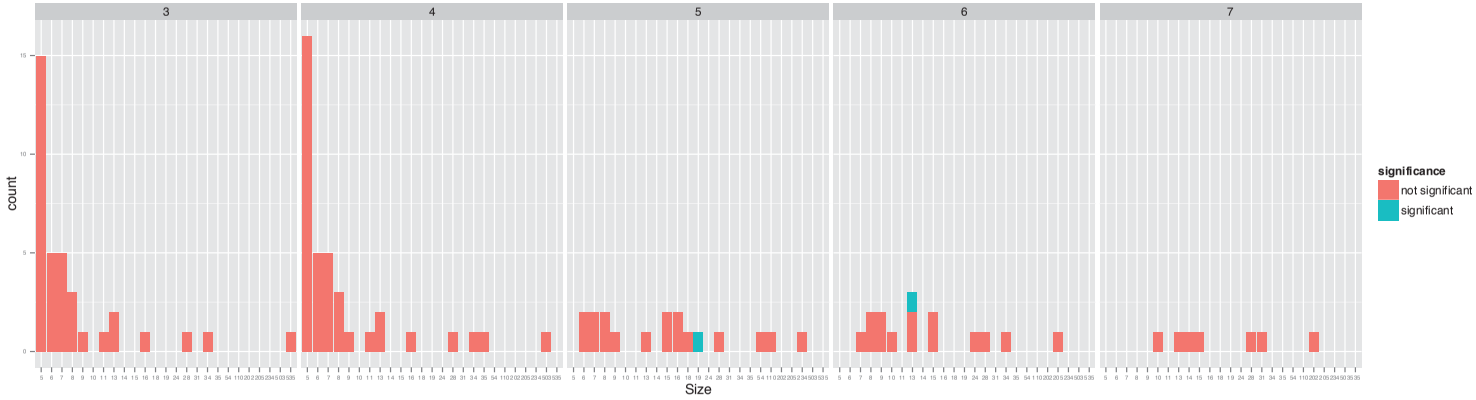
(d) α=0.80



The x-axis represents the size of k-pseudo cliques found at each k setting. The significant pseudo cliques are indicated in green. In Figures 5.1.3.4c and 5.1.3.4d at each k setting smallest k-pseudo cliques were the most numerous and the number of k-pseudo cliques discovered decreased as k-pseudo clique size increased.

Figure 5.1.3.4: k-pseudo clique size historgrams continued.

(e) α=0.75



The x-axis represents the size of k-pseudo cliques found at each k setting. The significant pseudo cliques are indicated in green. The maximum size of k-pseudo cliques discovered when $\alpha = 0.75$ was 535, around nine times as large as the k-pseudo clique of 61 vertices found at the other α settings.

Table 5.1.3.1: K-pseudo cliques with a lower than expected median MutSigCV p-value

k	α	Empirical p-value	Empirical q-value	Observed p-value	Permutation Median	Gene Set Genes
3	0.95,0.90,0.85	0.001	0.02	0.33	1	*CASP10, CASP8, FADD, FAS, FASLG, CFLAR, CASP3, PSEN2, TNFRSF10B, TNFSF10 TNFRSF10A*
	0.95,0.90, 0.85	8.0×10^{-3}	0.09	0.80	1	*CBL, EGFR, GRB2, PIK3R1, PTK2, SHC1, SRC, ERBB2, ERBB3, SH3KBP1 CRK, PTK2B, PLCG1, LYN, SYK, VAV1, CDH1, CTNNB1, CTNND1, APC AXIN1, GSK3B, CTNNA1, JUP, CDH2, CDH5, MUC1, ERBB4, NRG1, ABL1 NTRK1, GAB1, PIK3R2, BCAR1, NEDD9, PTPN11, JAK2, JAK1, PTPN6, PTPRM CBLB, IRS1, IGF1R, FLT4, KDR, FYN, INSR, LCK, PDGFRB, RET KHDRBS1, PTPN1, PTPRA, PRKCD, STAT3, STAT5A, BTK, LCP2, IRS2, YWHAG ZAP70*
4	0.95, 0.90 (k=4,5) ,0.85 (k=4,5)	3.0×10^{-3}	0.08	0.15	0.98	*EP300, SMAD2, SMAD3, SMAD4, PIAS4, SKI, SKIL*
	0.95, 0.90 (k=4,5)	6.0×10^{-3}	0.08	0.32	0.94	*CASP10, CASP8, FADD, FAS, FASLG, CFLAR*
	0.95, 0.90 (k=4,5), 0.85 (k=4,5)	0.01	0.08	0.78	1.00	*CBL, EGFR, GRB2, PIK3R1, PTK2, SHC1, SRC, ERBB2, ERBB3, SH3KBP1 CRK, PTK2B, PLCG1, LYN, SYK, VAV1, CTNNB1, MUC1, ERBB4 NRG1, ABL1, NTRK1, GAB1, PIK3R2, BCAR1, NEDD9, PTPN11, JAK2, JAK1 PTPN6, CBLB, IRS1, IGF1R, FYN, INSR, LCK, PDGFRB, RET, KHDRBS1 PTPN1, PTPRA, STAT5A, BTK, LCP2, ZAP70*
	0.95	0.01	0.08	0.21	0.97	*CDH1, CTNNB1, CTNND1, EGFR, PTPRM*
	0.95	0.01	0.09	0.32	1.00	*CASP8, FADD, TNFRSF10B, TNFSF10, TNFRSF10A*
5	0.95 (k=5,6)	4.0×10^{-3}	0.03	0.43	0.98	*CBL, EGFR, GRB2, PIK3R1, PTK2, SHC1, SRC, ERBB2, ERBB3, SH3KBP1, CRK, PTK2B, PLCG1, VAV1*
	0.8, 0.75	9.90×10^{-05}	1.98×10^{-3}	0.41	1.00	*CASP10, CASP8, CFLAR, FADD, FAS, FASLG, BID, PEA15, RIPK1, TNFRSF10B, TNFSF10, TRADD, VIL2, TNFRSF10A, CASP3, DEDD2, PSEN2, APP, VIM*
6	0.90 (k=6,7)	3.80×10^{-3}	0.01	0.43	0.98	*CBL, EGFR, GRB2, PIK3R1, PTK2, SHC1, SRC, ERBB2, ERBB3, CRK, PLCG1, PTK2B, SH3KBP1, VAV1*
	0.85	4.70×10^{-3}	0.04	0.70	1.00	*CBL, CRK, EGFR, GRB2, PIK3R1, PTK2, SHC1, SRC, PLCG1, ERBB2 ERBB3, INSR, PTK2B, SH3KBP1, VAV1, IRS1, CTNNB1, ERBB4, GAB1, PIK3R2 PTPN11, ABL1, FYN, KHDRBS1, LCK, PTPN1, RET, BCAR1, JAK2, SYK CBLB, NTRK1, ZAP70, IGF1R, PTPN6, LYN, BTK, LCP2*
	0.8, 0.75	6.00×10^{-4}	0.01	0.38	1.00	*CASP10, CASP8, CFLAR, FADD, FAS, FASLG, BID, PEA15, RIPK1, TNFRSF10B TNFSF10, TRADD, VIL2*
7	0.85	0.01	0.02	0.70	1.00	*CBL, CRK, EGFR, GRB2, PIK3R1, PTK2, SHC1, SRC, PLCG1, ERBB2, ERBB3, INSR, PTK2B, SH3KBP1, VAV1, IRS1, CTNNB1, ERBB4, GAB1, PIK3R2, PTPN11, ABL1, FYN, KHDRBS1, LCK, PTPN1, RET, BCAR1, JAK2, SYK*

k- pseudo cliques that were significant at more than one parameter setting are indicated in the α column. Red genes indicate those that were found to be significantly mutated in the univariate MutSigCV analysis.

K: The *k* parameter setting. **α**: the density threshold parameter setting. **Empirical p-value**: The raw permutation derived p-value. **Empirical q-value**: The Benjamini and Hochberg adjusted p-value**. Observed p-value**: The median MutSigCV p-value in the non-permuted network. **Permutation Median**: The median MutSigCV p-value for the k-pseudo clique across 10 000 permutations. **Gene Set Genes**: The genes that correspond to the significant k-pseudo clique for that row of the table.

SMAD4, *SMAD2*, *BRAF*, *TCF7L2*, and *CTNNB1*) were part of the KEGG colorectal cancer gene set (enrichment p-value of 2.34e-18).

Nine of the MutSigCV gene set; *KRAS*, *NRAS*, *TP53*, *APC*, *SMAD4*, *SMAD2*, *BRAF*, *TCF7L2*, and *CTNNB1*; were part of the KEGG pathways in cancer gene set, corresponding to a hypergeometric enrichment p-value of 1.35e-14. All of the k-pseudo clique analysis derived gene sets except for three: where $k = 4$ and α=0.80; $k = 5$ and α=0.80; and $k = 4$ and α=0.75 were more significantly enriched in the KEGG pathways in cancer gene set, than was the MutSigCV derived gene set. For seven k-pseudo clique derived gene sets a larger number of genes in the KEGG pathways in cancer gene set were discovered in comparison to the MutSigCV derived list (Table 5.1.3.2).

Of the eight significantly mutated k-pseudo cliques subjected to KEGG disease pathway analysis, five were most enriched in the 'ErbB signalling' pathway (Table5.1.5.1). The four ErbB family proteins (*EGFR*, *ERBB2*, *ERBB3*, and *ERBB4)* were present in all but one of these k-pseudocliques (where ErbB4 was absent). The three remaining k-pseudo cliques were highly enriched in the KEGG 'Apoptosis' pathway. These two k-pseudo cliques may represent two sets of samples with differing routes to cancer.

Table 5.1.3.2: KEGG pathway enrichment validation of MutSigCV and k-pseudo cliques analysis

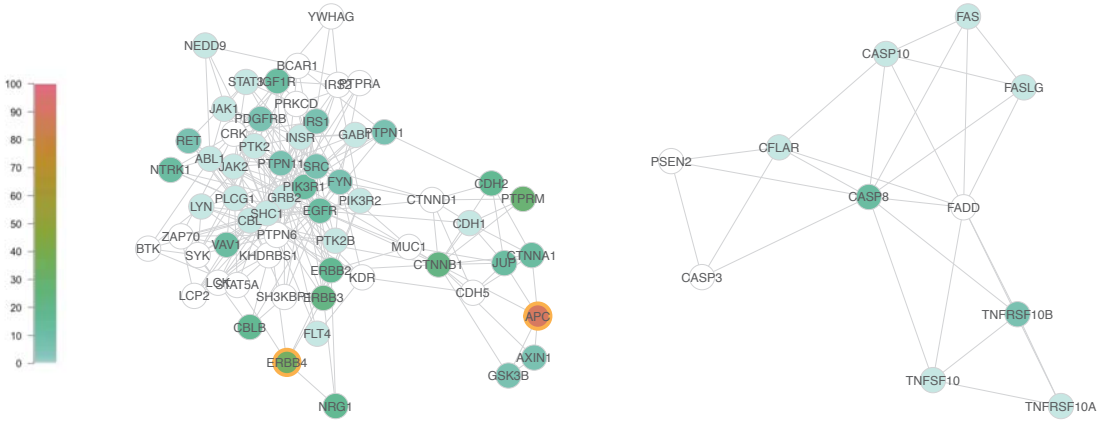
			Colorectal p-value		Pathways in Cancer p-value	
α	*k*	N proteins	N in pathway	network	N in pathway	network
0.95	3	72	7	2.22×10^{-9}	30	3.55×10^{-35}
	4	76	6	4.16×10^{-8}	28	1.22×10^{-33}
	5	14	N/A	N/A	8	1.13×10^{-11}
0.90	3 (same as 0.95 k=3)	72	7	2.22×10^{-9}	30	3.55×10^{-35}
	4	58	6	2.28×10^{-8}	27	2.08×10^{-33}
	5 (same as 0.90 k=4)	58	6	2.28×10^{-8}	27	2.08×10^{-33}
	6 (same as 0.95 k=5)	14	N/A	N/A	8	1.13×10^{-11}
	7 (same as 0.95 k=5)	14	N/A	N/A	8	1.13×10^{-11}
0.85	3 (same as 0.95, k=3)	72	7	2.22×10^{-09}	30	3.55×10^{-35}
	4	52	6	1.16×10^{-08}	23	5.84×10^{-28}
	5 (same as 0.85 k=4)	52	6	1.16×10^{-08}	23	5.84×10^{-28}
	6	38	N/A	N/A	15	9.92×10^{-18}
	7	30	N/A	N/A	12	1.62×10^{-14}
0.80	5	19	N/A	N/A	6	2.60×10^{-07}
	6	13	N/A	N/A	5	7.51×10^{-07}
0.75	5 (same as 0.80, k=5)	19	N/A	N/A	6	2.60×10^{-07}
	6 (same as 0.80, k=6)	13	N/A	N/A	5	7.51×10^{-07}

= more significantly enriched, or a larger number of genes intersecting the KEGG pathway than the MutSigCV derived gene list
= less significanlty enriched, or a smaller number of genes intersecting the KEGG pathway than the MutSigCV derived gene list

K-pseudo clique analysis settings that produced identical gene sets are indicated in the *k* column.

Figure 5.1.3.5: HuPPI α=0.95, 0.90, and 0.85 k=3 significantly mutated subnetworks.

(a) (b)



The vertex colours indicate the percentage of samples that carry at least one functional mutation in the gene. Orange outlines indicate genes mutated in at least 10% of samples. The gene lists for union of these subnetworks had the highest enrichment in both the KEGG colorectal cancer, and pathways in cancer gene sets. Figure 5.1.3.5b is contains *CASP8*, *CASP3*, *CASP10*, *FADD*, *FAS*, and *FASLG*, which are all part of the Apoptosis pathway.

5.1.4 Discussion

Using the k-pseudo cliques analysis I found no gene sets that were more highly enriched with known colorectal cancer genes than was achieved using the univariate MutSigCV test. However, in most cases the k-pseudo clique significant results were more enriched with known cancer genes than the results from the MutSigCV analysis (Table 5.1.3.2). This may indicate that additional mutations that are important for the progression of colorectal cancer occur in genes known to be important for progression in other cancers. Across individuals these mutations may occur at a frequency that is too low to be detected at a univariate level given the current sample sizes available to researchers.

5.1.4.1 Additional colorectal cancer associated genes in the α=0.95 and k=3 k-pseudo cliques

The publicly available KEGG disease pathways were last updated in November 2012. Since then, many comprehensive sequencing studies of various cancers have been completed. The KEGG colorectal cancer pathway gene list is a subset of the complete list of genes contributing to colorectal cancer progression. I wanted to understand if the k-pseudo cliques analysis had identified biologically plausible candidate genes for colorectal cancer progression that were not included in KEGG disease pathways at the time of the most recent update in November 2012. I obtained the list of 72 genes from significant k-pseudo cliques at the best performing parameter settings, as measured by KEGG pathway enrichment p-values, where $k = 3$ and $\alpha = 0.95$, and conducted a literature search. *CDH1*, the *CTNNB1* and *CTNNA1* complex, and *JUP* protein abundances were up-regulated in colorectal tumour cells in the TCGA colorectal cancer dataset (Zhang et al., 2014). *PIK3R1* and *PIK3R2*, which are isoforms of the *p85* regulatory subunit of the *PI3K* complex, were implicated by my method in colorectal cancer. Somatic mutation in these genes may lead to a reduction in the inhibitory effect of the p85 subunit on *PI3K* activation, ultimately resulting in cell proliferation in colorectal cancers (Papadatos-pastos et al., 2015). In colorectal cancer the

PTPRM promoter has been shown to be hyper-methylated in comparison to normal tissue (Laczmanska et al., 2012). *PTPRM* is a tumour suppressor, and its hyper-methylation may play an important role in colorectal cancer tumorigenesis. Methylation of *RET* has been associated with colorectal tumour progression from adenoma to cancer (Luo et al., 2013), and somatic mutation could also lead to *RET* inactivation (Sjöblom et al., 2006). *ERBB2* has been implicated as a druggable target in colorectal cancers (Muzny et al., 2012) and breast cancers (Koboldt et al., 2012; Slamon et al., 1987). Increased expression of *CBL* may play a role in the invasiveness properties of a subset of colorectal cancers (Cristóbal et al., 2014). *JAK1* signalling suppression increases apoptosis in *KRAS* mutant colorectal cancers (Kalimutho et al., 2012). The effects of these genes seem to apply to specific cancer subtypes and not across all colorectal cancers. These associations were discovered by the k-pseudo cliques analysis method and not by the univariate MutSigCV method. I have demonstrated how the k-pseudo cliques method identified additional genes in comparison to the univariate test, which are known to contribute to cancer. In addition these associations were not present in the KEGG colorectal cancer gene set, nor the KEGG pathways in cancer gene set.

5.1.4.2 Cancer related signalling pathways

The group of 11 genes identified by the k-pseudo cliques analysis at the density=0.95 and k=3 (Figure 5.1.3.5b) parameter settings seemed to represent the extrinsic *TRAIL* induced Apoptosis pathway (Ghavami et al., 2009), and 10 of these genes were part of the KEGG Apoptosis pathway. The k-pseudo clique containing 61 genes at α=0.95 and k=3 contained the ErbB family of receptors. MutSigCV (Lawrence et al., 2013) did not find any of the ErbB protein family to be significantly mutated. This finding may indicate genetic heterogeneity in colorectal cancer and the possibility of treating colorectal cancer using drugs and monoclonal antibodies that bind to ErbB receptors using a targeted approach.

5.1.4.3 Utility for complex disease genetics discovery

I used the KEGG colorectal cancer pathway enrichment analysis to compare the MutSigCV (Lawrence et al., 2013) method to identify significantly mutated genes in cancer to the network-based k-pseudo clique analysis method. I showed that the k-pseudo clique analyses in seven out of ten cases outperformed the MutSigCV analysis in KEGG pathways in cancer gene set enrichment, and the number of pathways in cancer genes recovered. The additional cancer candidate genes supplied by the k-pseudo clique methods may be false positives. However, given the findings that k-pseudo clique gene sets were enriched in colorectal cancer genes and the more recent literature linking significant k-pseudo clique gene sets to colorectal cancer, this method may be useful for uncovering new complex disease genetics by prioritising gene-level tests.

5.1.4.4 K-pseudo cliques analysis future improvements

The k-pseudo cliques method may perform even better with a more complete protein interaction network such as the irefweb (Turner et al., 2010), or humanNet (Lee et al., 2011). The k-pseudo clique method is limited by its reliance upon the DME program (Georgii et al., 2009). The time to enumerate all pseudo cliques within a graph, at any α, increases with the number of vertices and edges of the graph. The pseudo clique enumeration time also increases as α decreases (Georgii et al., 2009). K-pseudo clique analyses of large graphs may be extremely time consuming. The DME algorithm reverse search methodology means that for each dense module all possible single vertex extensions to the module are tested for satisfaction of the density criterion. This process can be parallelised (Georgii et al., 2009), but that is beyond the scope of this study. Parallelisation would decrease DME run-time, especially at lower α threshold, making this method more appropriate for analysis of large, and dense biological networks.

Setting the optimal parameters for the k-pseudo cliques analysis is not an easy task. For analysis of k-cliques (Palla et al., 2005), others decided that parameter settings for k-clique discovery in an unweighted network were optimal when the size of the second largest k-clique

was at least half the size of the largest k-clique (Palla et al., 2005). However, in the case of k-pseudo cliques analysis the intention was to reduce a single gene multiple testing burden while discovering well defined communities of functionally related genes. Any k-pseudo cliques that were found to be significantly mutated through permutation, but not at the gene level, have provided useful information regardless of their size. The functional interpretation of a k-pseudo clique is another matter, and may be easier for smaller k-pseudo cliques. In which case, choosing analysis parameters that do not result in a 'giant community' (Palla et al., 2005) may be preferred.

The k-pseudo clique analysis method is not limited to its use in analysis of cancer data, as is the MutSigCV tool. K-pseudo cliques analysis can be conducted using a dataset from which gene scores can be obtained and an appropriate biological network. This means that k-pseudo cliques analysis is appropriate for the analysis of: gene expression microarray data; GWAS data; RNA-seq; proteomic data; and metabolomic data. It could also be used to refine polygenic risk score models (Purcell et al., 2009), by identifying which genes with p-values in the range of 1×10^{-5}- 1×10^{-7} are the best potential disease gene candidates, thus improving the signal to noise ratio in polygenic risk models.

5.1.5 Conclusion

By using the k-pseudo cliques analysis method, the HuPPI2 PPI network to analyse the colorectal cancer Pan-Cancer dataset, and the KEGG disease pathways as gold standards for cancer pathway knowledge, I identified fewer genes known to contribute to colorectal cancer as the state of the art univariate method MutSigCV. However, the k-pseudo cliques analysis method outperformed the MutSigCV analysis for enrichment in genes from the KEGG pathways in cancer gene set at most parameter settings The k-pseudo clique analysis should now be conducted using a more comprehensive PPI network such as the irefweb network (Turner et al., 2010), or humanNet (Lee et al., 2011) in order to provide additional high quality prior biological information and demonstrate that the results are not network specific.

By integrating gene level test data with prior biological information in the form of a protein interaction network I was able to detect genes contributing to colorectal cancer subtypes that are part of the same, or closely related biological pathways. K-pseudo cliques analysis supplied additional candidate genes for colorectal cancer progression, which may be important for progression of undefined colorectal cancer subtypes. In future work I could investigate if mutations in particular subnetworks are important for discriminating between cancer subtypes. For instance, in endometrial carcinoma that *PTEN* mutations are associated with endometrioid tumours and that *TP53* mutations are associated with serous high-grade tumours, mutations in these genes being almost mutually exclusive.

The candidate disease genes suggested by the $k = 3$, and $\alpha = 0.95$ parameter settings were often implicated only in certain cancer subtypes. This implies that, in some cases, k-pseudo cliques represent a pathway that when subjected to perturbations in different genes across patients results in colorectal cancer, but subtly different types of colorectal cancer. One possible avenue for further research would be to investigate if there are PPI subnetworks within which proteins are either mutated in one cancer or another, but never both cancer types. This approach may be useful for analysis of heterogeneous diseases where many subtypes have been categorised using the same name. The k-pseudo clique method is generic

and can be used to analyse complex disease data of any modality provided that gene level scores can be obtained.

Table 5.1.5.1: KEGG enrichment of significantly mutated k-pseudo cliques larger than ten proteins.

α	k	N proteins	Pathway	Gene names intersecting with the pathway.	raw p-value	adjusted p-value
0.95, 0.9, 0.85	3	61	ErbB signaling pathway	PIK3R1, CBLB, ERBB2, PTK2, CBL, ERBB4 SRC, EGFR, ABL1, GRB2, PLCG1, SHC1 CRK, PIK3R2, GSK3B, NRG1, STAT5A, ERBB3, GAB1	3.25×10^{-31}	1.56×10^{-29}
0.95, 0.9, 0.85	3	11	Apoptosis	CASP3, TNFRSF10B, TNFSF10, FASLG, TNFRSF10A CASP8, CASP10, FADD, CFLAR, FAS	1.35×10^{-23}	9.45×10^{-23}
0.95, 0.90, 0.85	4 (0.90, and 0.85 k=5)	45	ErbB signaling pathway	PIK3R1, CBLB, ERBB2, PTK2, CBL, ERBB4 SRC, EGFR, ABL1, GRB2, PLCG1, SHC1 CRK, PIK3R2, STAT5A, NRG1, ERBB3, GAB1	5.78×10^{-32}	2.14×10^{-30}
0.95,0.90	(0.95 k=5), 6, (0.90 k=7)	14	ErbB signaling pathway	GRB2, PLCG1, SHC1, CRK, PIK3R1, ERBB2 ERBB3, PTK2, CBL, SRC, EGFR	1.69×10^{-24}	3.72×10^{-23}
0.85	6	38	ErbB signaling pathway	PIK3R1, CBLB, ERBB2, PTK2, CBL, ERBB4 SRC, EGFR, ABL1, GRB2, PLCG1, SHC1 CRK, PIK3R2, ERBB3, GAB1	6.28×10^{-29}	2.14×10^{-27}
0.85	7	30	ErbB signaling pathway	GRB2, ABL1, PLCG1, SHC1, CRK, PIK3R2 PIK3R1, ERBB2, ERBB3, PTK2, CBL, GAB1 ERBB4, SRC, EGFR	1.28×10^{-28}	4.03×10^{-27}
0.8	5	19	Apoptosis	CASP3, TNFRSF10B, TNFSF10, FASLG, RIPK1, TNFRSF10A CASP8, BID, TRADD, FADD, CASP10, FAS, CFLAR	5.48×10^{-28}	4.38×10^{-27}
0.8	6	13	Apoptosis	TNFRSF10B, TNFSF10, FASLG, RIPK1, CASP8, BID TRADD, FADD, CASP10, FAS, CFLAR	5.61×10^{-26}	4.49×10^{-25}

Chapter 5.2

Prioritising rheumatoid arthritis candidate disease genes using Region Growing Analysis.

5.2.1 Introduction

Genome-wide association studies (GWAS) aim to identify common genetic variants that distinguish a population of individuals (cases) that have a particular trait or disease from a population of healthy individuals who do not have the trait (controls) (McCarthy et al., 2008). Single nucleotide polymorphisms (SNPs) are loci of bi-allelic variation in the genome. The minor allele frequency for a SNP to be considered common is usually at least one percent. Cases and controls are genotyped at each SNP, the total number of which often exceeds 500 000. GWAS are conducted as a logistic regression (or more commonly a Cochran-Armitage trend test (Cochran, 1954; Armitage, 1955)) where for each genotyped SNP the number of copies of the minor allele are regressed onto the disease status (control/case) for all individuals. Single nucleotide polymorphism p-values are then corrected by the number of tests, resulting in the identification of the SNPs that predict disease status.

The SNPs associated with disease status are often located outside of genes and may be

correlated (in linkage disequilibrium (LD)) with the causal genes or SNPs.

Genome-wide association studies have been one of the most successful high-throughput genomic analysis methods of the past decade. They have identified thousands of common risk loci that contribute to a variety of complex diseases. However, the variants discovered so far typically explain a small proportion of the heritable variation of each complex disease. Explanations for this so-called *missing heritability* include rare genetic variants with large effect sizes that are likely to be too rare to pass GWAS quality control procedures (Manolio et al., 2009; McCarthy et al., 2008). An alternative explanation is that common variants of decreasingly small effect size account for the missing heritability of complex diseases as suggested for schizophrenia by Purcell et al. (2009). A combination of both is possible. The undiscovered common variants may have an effect size that is too small to be detected at the genome-wide significance level ($p < 5 \times 10^{-8}$). In order to address the sample size issue many studies have combined genome-wide samples across studies in so-called mega-analyses. The Wellcome Trust Case-Control Consortium (WTCCC1) (Burton et al., 2007) used the mega-analysis methodology to identify the common variants associated with seven major complex diseases including rheumatoid arthritis (RA). However, in all cases the total heritable genetic risk of each disease remained unexplained.

One reason that the total heritable risk was not explained may be that the GWAS significance threshold ($p < 5 \times 10^{-8}$) to account for multiple tests was too stringent. It is known that genes and their protein products do not perform their biological function alone in the cell. Proteins function as part of biochemical pathways. A malfunction to any one of the genes in a pathway may lead to the same disease. This scenario may describe the genetic heterogeneity of a complex disease, where differing genetic causes give rise to the same disease. This can explain why GWAS may not detect some of the SNPs associated with a disease. For each causal gene in a pathway only a subset of cases may carry the disease causing variant, with the remaining cases carrying variants in other genes in the pathway (Leiserson et al., 2013b). By using protein interaction data to identify pathways that are enriched for gene-level disease associations additional genes that contribute to the

disease may be discovered.

Protein interaction databases contain records of thousands of physical interactions between proteins. The physical interactions recorded in these databases can be represented as a graph where each protein in each interaction is represented by a vertex and each interaction between two proteins is represented by an edge. The graph representation of the protein interaction network enables graph theoretic analyses to be applied to the network to gain new insights into disease mechanisms and genetics.

5.2.1.1 Network-based analysis of GWAS data

The field of using network-based analyses for identification of genes contributing to complex diseases has been outlined in the introduction (1.11.4 on page 46). However, numerous approaches have been developed to prioritise candidate genes from GWAS data. The methods outlined below use an array of different PPI network analysis strategies and methods of deriving statistical significance of candidate genes or subnetworks.

The dmGWAS (Jia et al., 2011) method uses a seed and expand approach to identify 'modules' in a PPI network that are enriched for gene level GWAS associations. It iteratively adds genes to a module if it increases the module's association score above a threshold rate.

The Network Interface Miner for Multigenic Interactions (NiMMI) (Akula et al., 2011) first derives gene level association test statistics using VEGAS (Liu et al., 2010). It then defines small subnetworks of interacting protein pairs. Each subnetwork then includes proteins that are two edges away from one of the originating protein interaction pair. For each subnetwork, the genes are weighted using a modified google PageRank algorithm (Page et al., 1999). The gene weights are combined with the gene-level p-values, and a subnetwork Z-score is created. Using the NIMMI method to analysis Crohn's disease GWAS data Akula et al. (2011) identified Crohn's disease associated genes that were not found using single SNP, or gene level association tests.

The HYbrid SeT (HYST) (Li et al., 2012) analysis interrogates pairs of genes that are connected in a PPI network to identify associated gene pairs in which both genes were

not associated with the disease at the univariate level. In a Crohn's disease dataset, they successfully identified a triplet group of *JAK2*, *STAT3*, and *CCL2*. These three genes are part of the *JAK-STAT* pathway involved in Crohn's disease, and were not identified by SNP tests, or gene-level tests (Li et al., 2012).

Sharma et al. (2013) developed the GWAS-based Comorbidity Method (GCM) to explain some of the missing heritability of co-morbid complex diseases using molecular triangulation (Krauthammer et al., 2004), and *jActiveModules (Ideker et al., 2002)* to prioritise disease gene candidates for follow up studies. They discovered new associations between cholesterol levels and a SNP (rs234706, $p = 1.10^{-5}$) in the *CBS* gene.

The *jActiveModules* cytoscape application has also been used by Baranzini et al. (2009) to analyse Multiple Sclerosis (MS) GWAS data, where they confirmed links to immunological pathways and discovered potential links to neural pathways.

5.2.1.2 Study outline

I first used methods developed by Lehne et al. (2011) to derive gene level associations from SNP data. Then I used Region Growing Analysis (RGA) described in Lehne (2011) to analyse RA GWAS data from the WTCCC1 study (Burton et al., 2007) in order to prioritise the gene-level associations and provide additional candidate genes for RA.

Region Growing Analysis is a de-novo pathway discovery algorithm. RGA integrates gene-level test statistic data with PPI network data in order to identify sections of the PPI network that are enriched for a labelled set of genes. The non-labelled genes, by definition are additional candidate disease genes. A set of *seed* genes is first defined. The seed genes are labelled on a PPI network. From each seed gene a *region* is originated. The genes that are adjacent to a seed gene, or a path distance of two away from a seed gene are incorporated in to the region. This process continues until no additional genes can be included in the region. Regions that overlap are merged together. The significance of a region is determined depending on the size of the region. RGA is applied to 10 000 randomised networks and

the size of largest region discovered in each permutation is recorded to create an empirical distribution of the largest region size. A region discovered in the non-permuted network is significant if its size is larger than 95 percent of the largest regions discovered in the 10 000 randomised networks.

5.2.2 Methods

5.2.2.1 The WTCCC RA dataset

Data were downloaded from the Wellcome Trust Case Control Consortium (WTCCC) website in 2007. There were 1860 Rheumatoid Arthritis (RA) cases, and 2938 control samples retained after quality control procedures that rejected non-Caucasian ancestry, cryptic relatedness, sample duplication and contamination (Burton et al., 2007). The Affymetrix GeneChip 500k Mapping Array Set was used to genotype all samples. From the initial set of 500568 SNPs present on the Affymetrix GeneChip, 392575 SNPs were carried forward to association analysis. Reasons for SNP exclusion included a minor allele frequency (MAF) <0.01, missing genotyping or genotyping problems identified through genotyping cluster plots. Single nucleotide polymorphisms from the Major Histocompatibility complex (MHC) region were removed (chromosome 6, position 25,930,839 to position 33,495,825, NCBI GRCh37 assembly). The MHC region showed high association with RA status across its length (Burton et al., 2007). It is difficult to identify individual SNP association signals in this gene dense portion of chromosome 6, due to strong LD (correlation) between the SNPs.

5.2.2.2 Gene score generation

The method used to generate gene scores was developed by Lehne et al. (2011) and is described in this section. RGA requires a gene list as its input. In order to generate a ranked list of genes from GWAS data a single association score must be assigned to each gene. After WTCCC SNP quality control there remained more than 300 000 SNP tests. Firstly I assign each gene a score derived from the surrounding SNPs. Then SNPs were assigned to the

closest gene within a 40 kilobase window upstream, or downstream of the gene. More recent methods also take account of LD among SNPs when assigning SNPs to genes (Christoforou et al., 2012; Dudbridge & Koeleman, 2004; Conneely & Boehnke, 2007).

Once SNPs were assigned to genes, p-values were obtained for each gene based on the maximum chi-square value (χ_1^2) for any SNP assigned to the gene, the mean χ_1^2 of the SNPs assigned to the gene, or the median χ_1^2 from the top quartile of SNPs assigned to the gene. When calculated from the original dataset, whole gene test statistics were subject to gene length effects, the LD structure of SNPs, and number of SNPs per gene effects (Lehne et al., 2011). A longer gene is more likely to appear associated with the trait of interest when compared to a shorter gene because it is likely to contain more SNPs than a shorter gene (Lehne et al., 2011). In order to control for this problem, a permutation approach was chosen to obtain empirical p-values for the association of each gene with RA case status. Five hundred thousand case/control label permutations were carried out using PLINK (Purcell et al., 2007), giving an empirical distribution of 500 000 χ_1^2 values for each SNP. New gene wide p-values were calculated as the proportion of permuted datasets with χ_1^2 statistics higher than that of the original data. Three sets of gene-wide p-values were created; for the maximum χ_1^2 assigned to the gene (maxT); the median χ_1^2 of SNPs assigned to the gene (meanT); and the median p-value of the top quartile of χ_1^2 of SNPs assigned to the gene (topQ). These three sets of gene-wide p-values were used to create three sets of gene ranks: the smallest p-value being ranked first. The three gene rank sets were analysed separately using RGA.

5.2.2.3 Region Growing Algorithm

The Region Growing Algorithm was developed by Lehne (2011) and Christopher Tebbe, and is described in this section. From a set of gene ranks numerous ranked lists were generated to include the top ranked genes according to the threshold α, where α was set at 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 500, 1000, and 2000 genes. These gene lists contained the *seed genes*, all other genes were *non-seed genes*. Once genes are defined as seeds, they are

labelled on the PPI network (Figure 5.2.2.1a). The region growing algorithm then begins at all seed genes. RGA interrogates all of the adjacent vertices of the seed genes. If any adjacent vertices are seed genes, or adjacent to two or more seeds genes, then they are included in the *region* (Figures 5.2.2.1a, and 5.2.2.1b). This procedure is repeated for all vertices newly included in the region until no further vertices can be added (Figure 5.2.2.1c). Regions that intersect one another are fused to create a single larger region. When no additional genes can be included in any region, the number of vertices (genes) in each region is recorded as the *region size*.

The region growing algorithm finds regions composed of *seed genes*, and *non-seed genes* with at least two seed gene neighbours, the so-called outliers. The statistical significance of a region is achieved by using a degree-constrained label shuffling approach very similar to that which was used in Section 5.1.2.6.2 on page 161. For each permutation, the labels of all vertices are shuffled within the network and the region growing algorithm is run. The size of the largest region for each permutation is recorded. Labels are more likely to be swapped between vertices of similar degree than those of dissimilar degree. In each permuted network, the edges no longer represent the true physical interactions between the vertices. The permutation procedure removes the biological information about protein interactions encoded in the network by the edges. The process of degree-constrained label shuffling and region discovery in each permuted network was repeated 10 000 times.

Statistical evaluation of the regions was achieved through permutation testing. For each observed region, an empirical p-value was derived by recording the proportion of the largest regions found in the 10 000 randomised networks that were larger than the observed region. Observed regions with an empirical p-value <0.05 were interpreted as significantly large.

Figure 5.2.2.1: Visualising the steps in the RGA algorithm and its application to a protein interaction network.

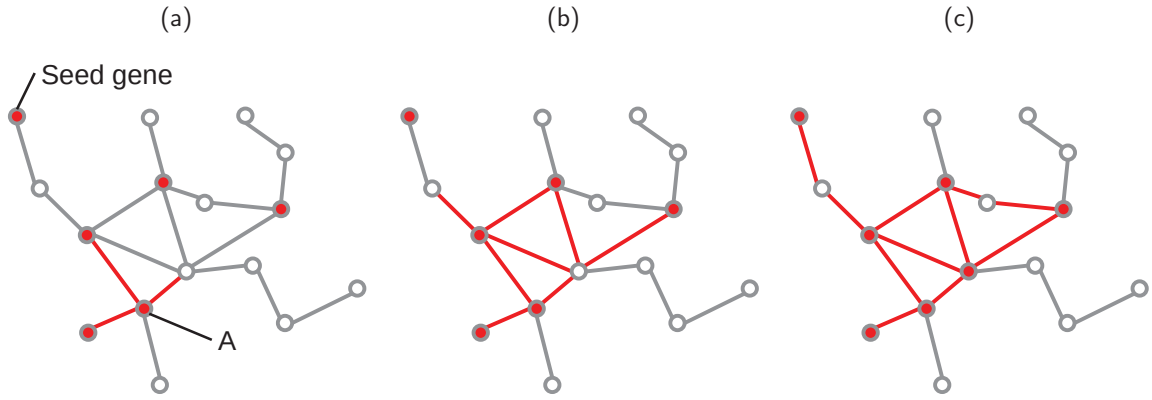


Figure 5.2.2.1a**:** Genes ranked within a user defined threshold α are *seed genes*. Starting from a seed gene A, RGA visits all neighbours. Neighbours which are seed genes, or are connected to two or more seed genes are included in the region. Figures 5.2.2.1b, and 5.2.2.1c: RGA is applied iteratively to all new genes added to the region. All seed genes originate their own region. Any intersecting regions are fused.

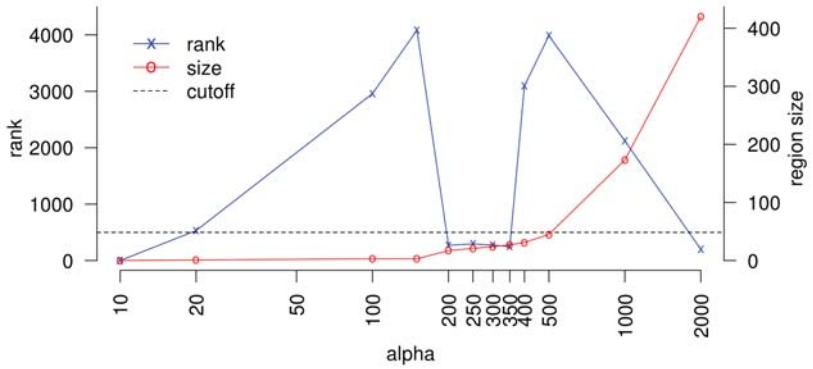
5.2.3 Results

Using the maxT gene ranks, there was a significant region found at $\alpha \in \{200, 300, 350\}$. The region was largest at α=350, and it contained 27 vertices (p=0.0247) (Figures 5.2.3.2a and 5.2.3.2b, and Table 5.2.6.1 on page 193). Using the topQ gene ranks, a significant region was identified at $\alpha \in \{30, 40, 50, 60, 70, 80, 90, 100, 200\}$ (Figures5.2.3.3a and 5.2.3.3b). The region increased in size up to α=200 where it consisted of 10 vertices (p=0.0403) (Table 5.2.6.2 on page 194). Using the meanT gene ranks, a significantly large region of six genes was identified at $\alpha \in \{60,70,80,90,150\}$. The region was composed of six vertices when RGA was applied to gene lists at α=60 through to α=150. The summary statistics for this region (Table 5.2.6.3 on page 194) were interpreted from the α threshold at which it attained the best rank (Figures 5.2.3.4a and Figure 5.2.3.4b). At α=60, the six gene region's p-value was p=7.70×10^{-03}.

The intersection of the significantly regions found using the maxT, meanT and topQ gene-wide p-value ranking strategies included six genes (Figure 2.5 and Table 2.4); *CSF2RB*,

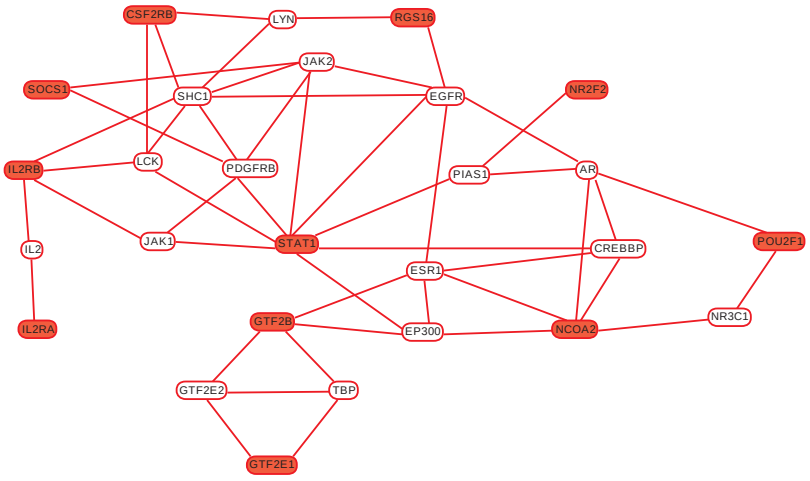
IL2, *IL2RB*, *IL2RA*, *LCK*, and *SHC1*.

Figure 5.2.3.1: Max T permutation results and significant regions



(a) Max T permutation results

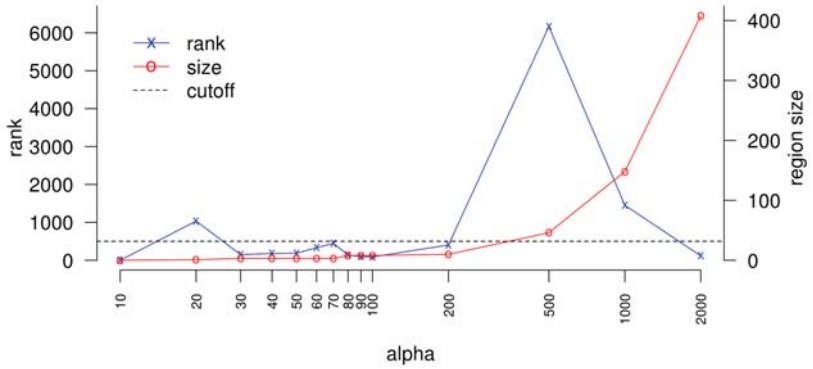
A highly ranked large region found using maxT ranks at α=350. The largest original dataset region size and rank against the number of top ranked genes designated as RGA seeds (alpha). Regions below the cut-off were within the 500 largest achieved by 10 000 network label shuffling permutations. The region at α=350 was ranked 247th and contained 27 genes.



(b) Max T region

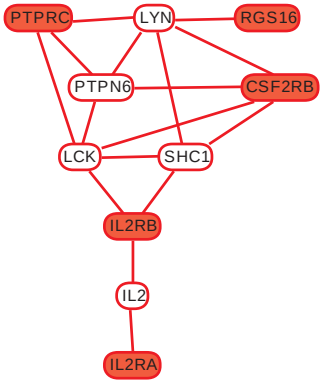
Rheumatoid Arthritis associated region in HuPPI2 using maxT ranks. The significant region of 27 genes found using maxT ranks at α=350. Physical interactions between gene products are represented by red edges. 'seed' genes (genes with ranks below the α threshold) were coloured red.

Figure 5.2.3.2: Top Q permutation results and significant region



(a) Top Q permutation results

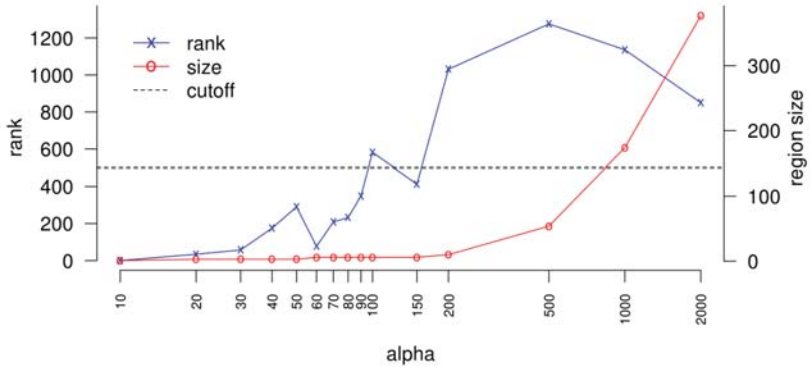
A highly ranked large region found using topQ ranks at α=200. The largest region rank and size against the number of top ranked genes designated as RGA seeds (alpha). Regions below the cut-off were within the 500 largest achieved by 10 000 network label shuffling permutations. The region at α=200 was ranked 403rd and contained 10 genes.



(b) Top Q region

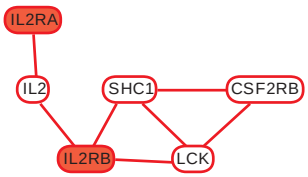
Rheumatoid Arthritis associated region in HuPPI2 using topQ ranks. The significant region of 10 genes found using topQ ranks at α=200. Physical interactions between gene products are represented by red edges. 'seed' genes (genes with ranks below the α) were coloured red.

Figure 5.2.3.3: Mean T permutation results and significant region



(a) Mean T permutation results

A highly ranked large region was found using meanT ranks at α= 60. The largest region size and rank against the number of top ranked genes designated as RGA seeds (alpha). Regions below the cut-off were within the 500 largest achieved by 10 000 network label shuffling permutations. The region at α=60 was ranked 77th and contained 6 genes.



(b) Mean T region

Rheumatoid Arthritis associated region in HuPPI2 using meanT ranks. The significant region of 6 genes found using meanT ranks at α=60. Physical interactions between gene products are represented by red edges. 'seed' genes (genes with ranks below the α threshold 60) were coloured red.

5.2.4 Discussion

The three different ranking strategies maxT, meanT and topQ cannot be interpreted as correct in isolation. All three strategies must be used in RGA and the results compared. Each ranking strategy assumes a genetic architecture. As I do not know the true genetic architecture of RA, the gene products found in significantly large region using the maxT, meanT and topQ gene-wide p-value ranking strategies are an approximation of the most promising candidate genes contributing to RA.

RGA has identified a subnetwork consisting of *IL2*, *IL2RA* and *IL2RB* that may contribute to RA. These genes were not associated with RA at the genome-wide significance level in the WTCCC GWAS (Burton et al., 2007), but associations with RA have been replicated at nominal p-values for association by Stahl et al. (2010) (*IL2* (rs6822844 4q27 p=7.00×10^{-04}), *IL2RA* (rs2104286 10p15, p=2.00×10^{-04}) and *IL2RB* (rs3218253 22q12 p=2.00×10^{-03})). *IL2RB*, and *IL2RA* are attractive candidate genes for RA because they encode proteins forming a complex (IL2R), a receptor required for the signal transduction of *IL2* (H. P. Kim, Imbert, & Leonard, 2006). The *IL2* / *IL2R* receptor system is involved in a number of immune processes including T-cell proliferation following antigen encounter, and regulatory T-cell function (H. P. Kim, Imbert, & Leonard, 2006). The three remaining genes in the subnetwork (*LCK*, *CSF2RB* and *SHC1*) were not found to be previously associated with RA by searching Entrez Gene (Maglott, 2004) and the Online Inheritance in Man (OMIM) (Hamosh et al., 2005) databases.

Region Growing Analysis is heavily reliant upon the quality and size of the network which is analysed. The HuPPI2 (Lehne & Schlitt, 2009) network represents less than 20 percent of all human protein coding genes. This means that genes involved in the disease process may be overlooked. As with many of the protein interaction meta-databases, HuPPI2 (Lehne & Schlitt, 2009) represents the union of physical protein interactions across all tissues. Some researchers have found that by using tissue specific networks disease gene prioritisation results can be improved (Jiang et al., 2009; Magger et al., 2012). In addition, the HuPPI2 network includes only protein interactions. Biochemical interactions such as transcription factor

binding may play an important role in complex diseases, but they are not represented in the HuPPI2 network. Networks such as Multinet (Khurana et al., 2013) that include protein-DNA physical interactions from the Encyclopdia of DNA Elements (ENCODE) (Gerstein et al., 2012) project will allow networks researchers to explore the role of protein interactions and gene expression regulation simultaneously.

Region Growing Analysis uses a unique network search strategy. By allowing an iterative search of gene neighbours two edges away from region members the discovery of regions composed of long chains of proteins is possible. This cannot be said for other methods that restrict the network search to two edges (NiMMI (Akula et al., 2011)), or gene pairs (HYST (Li et al., 2012)). However, The HyperModules software (Leung et al., 2014), takes a similar approach to Region Growing Analysis by labelling mutated genes in a PPI network as seed genes and then iteratively expanding the region by adding additional neighbouring genes if there is an increase in the correlation of the presence of a mutation in the module with a continuous, or categorical clinical variable. Leung et al. (2014) found 19 subnetworks in TCGA ovarian cancer exome sequence data that showed differential rates of survival in patients that carried mutations within the modules in comparison to those who did not. Ozgun et al. (2014) use a similar approach to identify subnetworks composed of mutually exclusively mutated genes in TCGA somatic mutation data.

5.2.5 Conclusions

Using GWAS and incorporating additional biological information in the form of a PPI network and the network based analysis RGA, I have identified a promising candidate region of six candidate RA genes, which included genes that had previously been found to be associated with RA at a nominal level using GWAS (Burton et al., 2007; Stahl et al., 2010). This may be an example of genetic heterogeneity where a mutation in any one of the members of the subnetwork confers additional RA risk.

Region growing analysis may be improved by considering more than one region in the observed network for statistical significance. This may allow the discovery of further plausible candidate genes. Additional candidate genes may also be discovered by applying RGA to tissue specific networks, or those which include multiple types of biochemical interactions.

5.2.6 Supplementary Materials

Table 5.2.6.1: Genes within the significant 27 gene region found using a rank threshold of α=350 based on maxT gene-wide empirical p-value ranks.

Gene	maxT rank	maxT p-value	Chro	Start	End
IL2RB	12	4.20×10^{-05}	22	20486973	20511039
IL2RA	21	3.06×10^{-47}	10	5979490	6030662
RGS16	44	2.75×10^{-03}	1	153804224	153810014
SOCS1	99	2.84×10^{-03}	16	11266017	11267782
POU2F1	172	0.03	1	138436084	138642676
NCOA2	190	0.01	8	66517566	66809329
STAT1	196	0.02	2	183693413	183738626
NR2F2	223	0.01	15	72999792	73013983
GTF2B	231	0.01	1	87435703	87474719
GTF2E1	291	0.14	3	117835641	117875998
CSF2RB	314	3.04×10^{-03}	22	20271284	20300485
EGFR	979	0.09	7	54924934	55112834
EP300	1829	0.23	22	24453007	24540433
IL2	3119	0.19	4	119099231	119104256
NR3C1	5767	0.37	5	137804733	137962312
LYN	6968	0.34	8	52260576	52392001
JAK2	7674	0.64	9	4940783	5083727
GTF2E2	8939	0.55	8	28980483	29060402
PIAS1	9014	0.56	15	45180222	45311150
ESR1	10236	0.71	6	149573851	149986188
TBP	12153	0.80	6	168368477	168387004
PDGFRB	12810	0.79	5	144641381	144683476
JAK1	13041	0.71	1	63407815	63541120
CREBBP	13888	0.94	16	3743994	3898607
LCK	14784	0.98	1	30833080	30867241
AR	NA	NA	X	60592006	60772474
SHC1	NA	NA	1	126296519	126308704

The table shows the gene symbol (gene);, The gene ranks according to maxT gene-wide empirical p-values (maxT rank); The gene-wide maxT empirical p-values (maxT p-value); The chromosome on which genes are located (chro), the start and end position of genes (start/end) and full gene names providing information about gene function (info).

Table 5.2.6.2: Genes within the significant region of 10 genes found using a rank threshold of α=200 and topQ gene-wide empirical p-value ranks.

gene	topQ rank	topQ p-value	Chro	start	end
IL2RB	12	4.20×10^{-05}	22	20486973	20511039
IL2RA	25	3.06×10^{-04}	10	5979490	6030662
RGS16	74	2.75×10^{-03}	1	153804224	153810014
CSF2RB	79	3.04×10^{-03}	22	20271284	20300485
PTPRC	195	9.42×10^{-03}	1	169775365	169893778
IL2	3112	0.19	4	119099231	119104256
LYN	5247	0.34	8	52260576	52392001
LCK	14791	0.98	1	30833080	30867241
PTPN6	NA	NA	12	6911844	6928943
SHC1	NA	NA	1	126296519	126308704

The table shows the gene symbol (gene); The gene ranks according to topQ gene-wide empirical p-values (topQ rank); The gene topQ gene-wide empirical p-values (topQ p-value); The chromosome on which genes are located (chro), The start and end position of genes (start/end) and full gene names providing information about gene function (info).

Table 5.2.6.3: Genes within the significant 6 gene region found using a rank threshold of α=60 based on meanT gene-wide empirical p-value ranks

gene	meanT rank	meanT p-value	chro	start	end
IL2RB	12	4.20×10^{-5}	22	20486973	20511039
IL2RA	25	3.06×10^{-04}	10	5979490	6030662
CSF2RB	79	3.04×10^{-03}	22	20271284	20300485
IL2	3112	0.19	4	119099231	119104256
LCK	14791	0.98	1	30833080	30867241
SHC1	NA	NA	1	126296519	126308704

The table shows the gene symbol (gene); The gene ranks according to topQ gene-wide empirical p-values (topQ rank); The gene topQ gene-wide empirical p-values (topQ p-value); The chromosome on which genes are located (chro), The start and end position of genes (start/end) and full gene names providing information about gene function (info).

Chapter 6

Conclusions

6.1 Overview of the thesis

Chapter 1 introduced the topic of research undertaken in this thesis. First an introduction to cancer exome sequencing and TCGA and Pan-Cancer projects were given, followed by an overview of machine learning methods in the context of cancer gene discovery. An introduction to graph theory and protein interaction networks was given, and also discussed in the context of prioritisation of complex disease gene candidates. An outline of the key research goals was then given.

Chapter 2 described the exploratory analyses conducted on TCGA colorectal cancer exome sequence data from June 2012, December 2012, and from the Pan-Cancer analysis project. Non-metric multidimensional scaling and PCA were used to explore the technical effects in the three datasets. A sequence technology effect was discovered in TCGA datasets that was reduced in the Pan-Cancer dataset. This led to the decision to use the Pan-Cancer exome sequence data for all further cancer analysis undertaken in the thesis.

Chapter 3 explored the problem of discovering mutations that correlate with cancer grade and stage across cancer types. Logistic regression models were created to identify the genes carrying functional mutations that were associated with low grade and high grade tumours (the same was done for low stage and high stage tumours). This was done across, and within three types of adenocarcinoma (ovarian adenocarcinoma, renal cell carcinoma, and endometrial carcinoma). Functional mutations in *TP53* were associated with a high grade status across adenocarcinomas. In endometrial carcinoma previously known associations between *TP53* mutations and high grade status, and *PTEN* mutations and low grade status (Garcia-Dios et al., 2013) were confirmed. Across and within adenocarcinomas there were no mutated genes associated with cancer stage when adjusting for the covariates age, gender, grade, and tumour type.

Chapter 4 applied a Random Forest approach to predict the class membership of Pan-Cancer samples using exome sequence data. In Chapter 4.1 Pan-Cancer samples were assigned to one of five high-order cancer types (adenocarcinoma, squamous cell carcinoma, urothelial carcinoma, a brain cancer, and a blood cancer) and mutated genes that discriminated

each cancer type from all others were identified. The genes that were most discriminatory between cancers were those that were mutated almost exclusively in a single cancer type (Kandoth et al., 2013a) including *VHL* (renal clear cell adenocarcinoma), *APC* (colorectal adenocarcinoma), and *DNMT3A* (acute myeloid leukaemia).

Chapter 4.2 developed the work from Chapter 4.1 and investigated the feasibility of creating a tool to predict the origin of metastatic cancers of unknown primary origin (CUP) in order to help guide the treatment choices of clinicians. Using the Pan-Cancer somatic mutation data, Random Forest models were created which assigned Pan-Cancer primary tumours of known origin to their tissue with around 70 percent accuracy.

Chapter 5 tackled the problem of candidate disease gene prioritisation for complex diseases by using network based approaches. Chapter 5.1 described the development of *k-pseudo cliques analysis*: a network-based method to prioritise disease gene candidates by integrating gene-level test statistics (from MutSigCV (Lawrence et al., 2013)) with PPI network data. K-pseudo cliques analysis finds communities of highly interacting proteins, called *pseudo-cliques*, in a PPI network using the DME pseudo clique enumeration software (Georgii et al., 2009; Tsuda & Georgii, 2009). Pseudo-cliques that overlap by $k-1$ proteins are grouped to give a set of k-pseudo cliques. The k-pseudo cliques are tested for enrichment with low p-values from gene level test statistics using a vertex label re-assignment permutation test. K-pseudo cliques that contained lower than expected MutSigCV test statistics were discovered. The 'significant' k-pseudo cliques composed of at least ten proteins were enriched for either the KEGG Apoptotsis pathway, or ErbB signalling pathway (Kanehisa & Goto, 2000; Kanehisa et al., 2014).

Chapter 5.2 described an investigation into candidate disease gene prioritisation by integrating gene-level genetic association data derived from a genome wide association study (GWAS) and PPI network data. *Region Growing Analysis* (RGA) (Lehne, 2011) was used to demonstrate that genes associated with rheumatoid arthritis (RA) are clustered in the HuPPI2 network (Lehne & Schlitt, 2009). A significant subnetwork of six genes was discovered. It included three genes known to be associated with rheumatoid arthritis (*IL2RA*,

IL2, and *IL2RB*) and three additional RA candidate disease genes.

6.2 Research aims revisited

At the end of Chapter 1 three research aims were outlined. These are now revisited to ascertain whether they have been fulfilled.

- Identify mutated genes associated with grade, and stage across cancers

This was addressed in Chapter 3. Cancer grade was dichotomised in to *low* and *high* categories to mitigate inter-pathologist variation in tumour grading. Cancer stage was dichotomised in the same way (*low stage* / *high stage)* to distinguish cancers that were confined to the primary tissue and those that had spread beyond. Logistic regression was used to find that *TP53* mutations are associated with high grade status across three types of adenocarcinoma (ovarian adenocarcinoma, renal cell carcinoma, and endometrial carcinoma). No mutated genes were associated with cancer stage when adjusted for age, gender, cancer grade and cancer type. The aim was met, however other approaches may have more success.

- Identify the mutated genes that discriminate five high-order cancer types (adenocarcinomas, squamous cell carcinomas, urothelial carcinomas, blood, and brain cancers) from one another

This aim was tackled in Chapter 4.1. The route taken to identify the mutated genes that discriminated the five cancer types from one another was to use ten pairwise Random Forest analyses. The Random Forest is designed as a classification algorithm, but it also simultaneously performs feature selection, to identify the set of features which best discriminate between the two cancer types included in each comparison. For each comparison the data was divided into a two thirds training set for model building and one third test set for model evaluation. The majority class in the training set was down-sampled to be equal to the minority class. This was because of the severe class imbalances across cancer types. It was found to be most severe between urothelial carcinoma (89 samples) and

adenocarcinoma (791 samples) cancer types. All Random Forest models performed with classification accuracy above 0.7 and included mutated proteins. However, the biological interpretation of how mutations among the proteins interact to discriminate the cancer types needs to be investigated further.

The aim to produce a tool to discriminate between cancers of different tissue types based on whole exome sequence data was achieved in Chapter 4.2. The approach used protein features, variant frequency features, and transition and transversion frequency features to discriminate between the ten Pan-Cancer solid tumour types. This time a ten class multi-class Random Forest was used to discriminate between the ten cancer classes and identify the features that discriminated between the cancers. Two thirds of the dataset was used for model building (the training set) and one third was used for model evaluation (the test set). Two sampling schemes were used to balance the cancer classes in the training set. The first involved downsampling classes to be equal to the minority class (bladder urothelial carcinoma). The second strategy used a combination of up-sampling and down-sampling to create a larger training set. The Random Forest models based on the two sampling schemes performed similarly. Most of the variant frequency, transition and transversion frequency features were included in the classification models indicating that mutation frequency is a good discriminator between cancer types. Genes that are exclusively mutated in certain cancer types were also included in the models including *VHL* (renal clear cell carcinoma), *APC*, and *KRAS* (colorectal adenocarcinoma).

- Use prior biological information in the form of a protein interaction network to suggest new complex disease gene candidates

The final research aim was addressed in Chapter 5. K-pseudo cliques analysis was performed in Chapter 5.1. It prioritised candidate disease genes by integrating PPI network data with gene level test statistics. Network-based methods have been used successfully to discover new disease genes in complex diseases (Akula et al., 2011; Li et al., 2012) and cancer (Chuang et al., 2007; Cerami et al., 2010; Hofree et al., 2013). However, k-pseudo cliques analysis is the only method to enumerate all pseudo cliques in a PPI network, identify overlapping

pseudo cliques known as k-pseudo cliques, and identify which of the k-pseudo cliques are enriched for low p-values from gene-level tests. Firstly, pseudo cliques were identified using the DME software (Georgii et al., 2009) at densities (α) ranging from 0.95 to 0.75 in descending steps of 0.05. Then the clique percolation approach used in CFinder (Palla et al., 2005, 2007) was used to identify k-pseudo cliques: pseudo cliques that overlapped by $k-1$ genes. MutSigCV (Lawrence et al., 2013) was used to measure the extent to which each gene was significantly mutated in the Pan-Cancer colorectal cancer exome sequence data. MutSigCV p-values were assigned to each gene in the PPI network. The median p-value of the genes was computed for each k-pseudo clique. A vertex label reassignment procedure was used to create 10000 permuted networks where the median p-value of each k-pseudo clique was recorded. An empirical p-value was assigned to each k-pseudo clique in the observed network. This corresponded to its rank among the permuted scores divided by 10000. A k-pseudo clique was deemed to be significant at an FDR<0.1 (Benjamini & Hochberg, 1995).

Twelve significant k-pseudo cliques were discovered that in total contained 87 genes. Lists of genes derived from the significant k-pseudo cliques at each parameter setting often outperformed the MutSigCV results in terms of enrichment in the KEGG 'pathways in cancer' gene set. The k-pseudo cliques settings where α=0.95 and k=3 performed best. Two significant k-pseudo cliques were discovered at α=0.95 and k=3. One of which contained all four ErbB signalling proteins (*EGFR*, *ERBB2*, *ERBB3*, and *ERBB4*) which were not found to be significantly mutated by MutSigCV.

Region Growing Analysis was used to prioritise gene-level associations derived from GWAS data in the WTCCC rheumatoid arthritis dataset (Burton et al., 2007) using prior biological information in the form of the HuPPI2 (Lehne & Schlitt, 2009) PPI network. Region Growing Analysis confirmed a subnetwork of interacting proteins (*IL2RA*, *IL2RB*, *IL2*) that were known to contribute to the disease. Association p-values for single nucleotide polymorphisms close to these three genes did not achieve genome-wide significance in the WTCCC GWAS (Burton et al., 2007). Three other genes (*LCK*, *CSF2RB*, and *SHC1*) were included in the

subnetwork that had not previously been associated with rheumatoid arthritis. The RGA method therefore successfully prioritised disease genes and suggested additional candidate genes by integrating gene-level test statistics with PPI network data.

6.3 Contributions

The key contributions of this thesis are listed below.

- The association between *TP53* functional mutations and high cancer grade across three types of adenocarcinoma (ovarian adenocarcinoma, renal cell carcinoma, and endometrial carcinoma) when adjusted for cancer type, age, gender, and stage.
- A Random Forest model that can be used to predict the tissue of origin of cancers of unknown primary origin using whole exome sequence data.
- A method to prioritise candidate disease genes (k-pseudo cliques analysis) based on gene level test statistic data and protein interaction network data.

6.4 Limitations

In Chapter 3 TCGA tumour grade was treated as the outcome, but grading is a subjective measure with significant inter-pathologist variation (Al-Aynati et al., 2003; Engers, 2007). It was not possible to account for inter-pathologist variation in grading because the pathology reports for each tumour were censored. Even after dichotomising the cancer grade in to low and high categories inter-pathologist variation in tumour grading may have prevented some of the features that associated with tumour grade from being identified.

Chapter 4.1 identified sets of genes that discriminate between five cancer types. However, the interpretation of how these mutated genes interact to discriminate between the cancer types is not easily understood when up to 100 proteins were included in some models.

The Random Forest models developed in Chapter 4.2 to assign CUP to their tissue of origin based on exome sequence data have limitations. The Pan-Cancer data that was used

to create the models is not a complete set of cancer types. Additional data are needed to validate these results in CUP.

The k-pseudo cliques analysis developed in Chapter 5.1 is limited by its reliance upon the Dense Module Enumeration (Georgii et al., 2009; Tsuda & Georgii, 2009) (DME) algorithm to find the pseudo-cliques. DME Computation time increases as the density and size of the input network increases. This may prohibit the analysis of larger networks.

The Region Growing Analysis that was used in Chapter 5.2 uses the size of the largest region in the observed network ranked among the sizes of the largest regions achieved in each of 10000 permuted networks to derive statistical significance. Therefore only a single region can achieve statistical significance.

6.5 Future work

Future work will involve the extension and improvement of analyses and methodologies presented in this thesis. Chapter 3 investigated the mutated genes associated with cancer grade in the Pan-Cancer dataset. However, tumour grading is subject to significant inter-pathologist variation (Al-Aynati et al., 2003; Engers, 2007). Work has begun to account for this variation by using the grade annotations of a single expert pathologist (Dr Salvador Diaz-Cano) based on pathology slides available from the cancer digital slide archive (Gutman et al., 2013) instead of TCGA grade annotations. The same logistic regression analyses will then be conducted.

The Random Forest model to assign a tissue of origin to CUPs must be validated using a set of CUPs that have undergone whole exome sequencing. Genomics England is undertaking the genome sequencing of cancers of unknown primary which can be used to validate the Random Forest models when the data becomes available. The exome sequence analysis of 21 TCGA cancer types has recently been standardised (Lawrence et al., 2014). The data could be used to create a more comprehensive prediction model.

The k-pseudo cliques analysis developed in Chapter 5.1 requires a more scalable approach to pseudo clique discovery if it is to be applied to large PPI networks. A further development

may be to sacrifice the solution to the pseudo-clique enumeration problem that is achieved by DME (Tsuda & Georgii, 2009), for a heuristic approach which may be faster for dense networks.

The Region Growing Analysis that was used in Chapter 5.2 may be improved by considering more than one region for statistical significance. A more appropriate approach may be to rank all regions discovered across 10000 permutations, and consider any region from the non-permuted network as significant if it is larger than the *n*th largest region achieved by network permutation.

An extension to the methods applied in Chapters 3 and 5 to identify subnetworks within a PPI network which exhibit association with a binary outcome could be developed. A set of *seed genes* is established using logistic regression to identify genes that are significantly associated with the outcome once adjusted for relevant covariates. The seed genes are mapped on the PPI network. Starting from a seed gene the logistic regression analysis is repeated after adding an adjacent gene to the model. AIC (Akaike, 1974) is used to decide whether or not to include the adjacent gene in the subnetwork. In this way, all genes adjacent to the subnetwork members are tested. The analysis is repeated until no further genes can be added to the model. Such a method would identify subnetworks of mutated genes that are independently associated with an outcome.

6.6 Concluding remarks

This thesis has contributed to the understanding of the molecular correlates of cancer grade across multiple cancers. It has identified sets of genes which potentially discriminate between high-order cancer types, which in turn led to the development of a tool which could be used to annotate CUP with a tissue of origin. In the future this may aid patient care. The k-pseudo cliques analysis developed in this thesis has application beyond cancer and may be used to prioritise candidate genes for any complex disease.

The further development of biomarkers and clinical cancer tests may require more comprehensive phenotypic information, as phenotypic information is usually more informative

than genomic information for clinical classification tasks. Projects like Genomics England should provide an excellent basis for the development of such tests through the integration of electronic health records with patient genomic data.

Gene-level analyses of exome sequence data in cancer have been successful in identifying some of the mutations that cause cancer. However, the comprehensive set of molecular alterations that cause cancer and contribute to cancer progression is likely to include other types of mutation. Whole genome sequencing may identify mutations in transcription factor binding sites, and methylation data may indicate genes which have been silenced in cancer. Even by integrating the analyses of these different data types, genetic heterogeneity may prevent the discovery of causative mutations using univariate tests. Where genetic heterogeneity is present in integrated datasets, network-based approaches have the potential to disentangle some of that heterogeneity and improve our understanding of the causes of cancer, and other complex diseases.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Akula, N., Baranova, A., Seto, D., et al. (2011). A network-based approach to prioritize results from genome-wide association studies. *PLoS ONE*, 6(9).

Al-Aynati, M., Chen, V., Salama, S., et al. (2003). Interobserver and intraobserver variability using the Fuhrman grading system for renal cell carcinoma. *Archives of Pathology and Laboratory Medicine*, 127(5), 593–596.

Albert, R. & Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1), 47–97.

Alcaraz, N., Kücük, H., Weile, J., Wipat, A., & Baumbach, J. (2011). KeyPathwayMiner: Detecting Case-Specific Biological Pathways Using Expression Data. *Internet Mathematics*, 7(4), 299–313.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., et al. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463), 415–21.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3), 375–386.

Bader, G. D., Betel, D., & Hogue, C. W. V. (2003). BIND: The Biomolecular Interaction Network Database. *Nucleic Acids Research*, 31(1), 248–250.

Banerji, S., Cibulskis, K., Rangel-Escareno, C., et al. (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486(7403), 405–409.

Barandela, R. & Sã, J. S. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3), 849–851.

Baranzini, S. E., Galwey, N. W., Wang, J., et al. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human Molecular Genetics*, 18(11), 2078–2090.

Barrenas, F., Chavali, S., Holme, P., Mobini, R., & Benson, M. (2009). Network Properties of Complex Human Disease Genes Identified through Genome-Wide Association Studies. *PLoS ONE*, 4(11), 2–7.

Beggs, A. D., Jones, A., El-Bahwary, M., et al. (2013). Whole-genome methylation analysis of benign and malignant colorectal tumours. *Journal of Pathology*, 229(5), 697–704.

Bell, D., Berchuck, A., Birrer, M., et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353), 609–615.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), 53–59.

Berglund, E. C., Kiialainen, A., & Syvänen, A.-C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*, 2(1), 23.

Bocker, T., Diermann, J., Friedl, W., et al. (1997). Microsatellite instability analysis: a multicenter study for reliability and quality control. *Cancer research*, 57(21), 4739–4743.

Boland, C. R., Thibodeau, S. N., Hamilton, S. R., et al. (1998). A National Cancer Institute workshop on microsatellite instability for cancer detection and familial predisposition: Development of international criteria for the determination of microsatellite instability in colorectal cancer. In *Cancer Research*, volume 58 (pp. 5248–5257).

Boveri, T. (2008). Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of cell science*, 121 Suppl(Supplement 1), 1–84.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Burton, P. R., Clayton, D. G., Cardon, L. R., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678.

Caulfield, M., Davies, J., Dennys, M., et al. (2015). *The 100,000 Genomes Project Protocol*. Technical Report Retrieved from http://www.genomicsengland.co.uk/library-and-resources/.

Cerami, E., Demir, E., Schultz, N., Taylor, B. S., & Sander, C. (2010). Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE*, 5(2).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, P. W. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Christoforou, A., Dondrup, M., Mattingsdal, M., et al. (2012). Linkage-disequilibrium-based binning affects the interpretation of GWASs. *American Journal of Human Genetics*, 90(4), 727–733.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(140), 140.

Cibulskis, K., Lawrence, M. S., Carter, S. L., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3), 213–9.

Ciriello, G., Cerami, E., Sander, C., & Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2), 398–406.

Ciriello, G., Miller, M. L., Aksoy, B. A., et al. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10), 1127–1133.

Cochran, W. G. (1954). Some Methods for Strengthening the Common Chi-square Tests. *Biometrics*, 10(4), 417.

Conneely, K. N. & Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *American journal of human genetics*, 81(6), 1158–1168.

Cowley, M. J., Pinese, M., Kassahn, K. S., et al. (2012). PINA v2.0: Mining interactome modules. *Nucleic Acids Research*, 40(D1), 862–865.

Creighton, C. J., Morgan, M., Gunaratne, P. H., et al. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456), 43–49.

Cristóbal, I., Manso, R., Rincón, R., et al. (2014). Up-regulation of c-Cbl suggests its potential role as oncogene in primary colorectal cancer. *International Journal of Colorectal Disease*, 29, 641–641.

Csárdi, G. & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, Complex Sy, 1695.

Das, J., Gayvert, K. M., Bunea, F., Wegkamp, M. H., & Yu, H. (2015). ENCAPP: elastic-net-based prognosis prediction and biomarker discovery for human cancers. *BMC Genomics*, 16(1), 1–13.

Dawson, S.-J., Tsui, D. W. Y., Murtaza, M., et al. (2013). Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer. *N Engl J Med*, 368(13), 1199–1209.

de la Chapelle, A. & Hampel, H. (2010). Clinical relevance of microsatellite instability in colorectal cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(20), 3380–7.

Dees, N. D., Zhang, Q., Kandoth, C., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome research*, 22(8), 1589–98.

Dennis, J. L. & Oien, K. a. (2005). Hunting the primary: Novel strategies for defining the origin of tumours. *Journal of Pathology*, 205(2), 236–247.

Diaz, L. a. & Bardelli, A. (2014). Liquid biopsies: Genotyping circulating tumor DNA. *Journal of Clinical Oncology*, 32(6), 579–586.

Dietmaier, W., Wallinger, S., Bocker, T., et al. (1997). Diagnostic microsatellite instability: Definition and correlation with mismatch repair protein expression. *Cancer Research*, 57(21), 4749–4756.

Ding, L., Kim, M., Kanchi, K. L., et al. (2014). Clonal Architectures and Driver Mutations in Metastatic Melanomas. *PLoS ONE*, 9(11), e111153.

Dorigo, M., Di Caro, G., & Gambardella, L. M. (1999). Ant algorithms for discrete optimization. *Artificial life*, 5(2), 137–172.

Dudbridge, F. & Koeleman, B. P. C. (2004). Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *American journal of human genetics*, 75(3), 424–435.

Edge, S. B. & Compton, C. C. (2010). The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM.

Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1), 36–48.

Efron, B. & Tibshirani, R. (1985). THE BOOTSTRAP METHOD FOR ASSESSING STATISTICAL ACCURACY. *Behaviormetrika*, 12(17), 1–35.

Engers, R. (2007). Reproducibility and reliability of tumor grading in urological neoplasms. *World journal of urology*, 25(6), 595–605.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7), 1575–1584.

Epstein, J. I. (2010). An update of the Gleason grading system. *The Journal of urology*, 183(2), 433–40.

Fidalgo, P. O., Cravo, M. L., & Nobre-Leitão, C. (1998). Re: A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: meeting highlights and Bethesda Guidelines. *Journal of the National Cancer Institute*, 90(12), 939–940.

Flicek, P., Amode, M. R., Barrell, D., et al. (2014). Ensembl 2014. *Nucleic Acids Research*, 42(D1), 749–755.

Frampton, G. M., Fichtenholtz, A., Otto, G. a., et al. (2013). Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature biotechnology*, 31(11), 1023–31.

Franceschini, A., Szklarczyk, D., Frankild, S., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(Database issue), D808–15.

Fuhrman, Susan A and Lasky, Larry C and Limas, C. (1982). Prognostic significance of morphologic parameters in renal cell carcinoma. *The American journal of surgical pathology*, 6(7), 655—-664.

Gandhi, T. K. B., Zhong, J., Mathivanan, S., et al. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature genetics*, 38(3), 285–293.

Garcia-Dios, D. a., Lambrechts, D., Coenegrachts, L., et al. (2013). High-throughput interrogation of PIK3CA, PTEN, KRAS, FBXW7 and TP53 mutations in primary endometrial carcinoma. *Gynecologic oncology*, 128(2), 327–34.

Garraway, L. a. & Lander, E. S. (2013). Lessons from the cancer genome. *Cell*, 153(1), 17–37.

Gavin, A.-C., Bösche, M., Krause, R., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868), 141–147.

Genomics England (2015). *A Framework for Industry Engagement: Genomics Enterprises Prospectus*. Technical Report Retrieved from http://www.genomicsengland.co.uk/library-and-resources/.

Georgii, E., Dietmann, S., Uno, T., Pagel, P., & Tsuda, K. (2009). Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics*, 25(7), 933–940.

Gerstein, M. B., Kundaje, A., Hariharan, M., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414), 91–100.

Ghavami, S., Hashemi, M., Ande, S. R., et al. (2009). Apoptosis and cancer: mutations within caspase genes. *Journal of medical genetics*, 46, 497–510.

Girvan, M. & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12), 7821–6.

Goh, K.-I., Cusick, M. E., Valle, D., et al. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21), 8685–8690.

Govindan, R., Ding, L., Griffith, M., et al. (2012). Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*, 150(6), 1121–1134.

Gutman, D. A., Cobb, J., Somanna, D., et al. (2013). Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *Journal of the American Medical Informatics Association*, 20(6), 1091–1098.

Gymrek, M., Golan, D., Rosset, S., & Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, 22(6), 1154–1162.

Hammond, M. E. H., Fitzgibbons, P. L., Compton, C. C., et al. (2000). College of American Pathologists Conference XXXV: Solid tumor prognostic factors - Which, how and so what? Summary document and recommendations for implementation. *Archives of Pathology and Laboratory Medicine*, 124, 958–965.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. a., & McKusick, V. a. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(DATABASE ISS.), 514–517.

Han, G., Sidhu, D., Duggan, M. a., et al. (2013). Reproducibility of histological cell type in high-grade endometrial carcinoma. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 26(12), 1594–604.

Hanley, J. A. & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3), 839–843.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., et al. (2004). IntAct: an open source molecular interaction database. *Nucleic acids research*, 32(Database issue), D452–D455.

Highnam, G., Franck, C., Martin, A., et al. (2013). Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Research*, 41(10), 1–7.

Hoadley, K. a., Yau, C., Wolf, D. M., et al. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, 158(4), 929–944.

Hofree, M., Shen, J. P., Carter, H., Gross, A., & Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature methods*, 10(11), 1108–15.

Höglund, M., Gisselsson, D., Säll, T., & Mitelman, F. (2002). Coping with complexity: Multivariate analysis of tumor karyotypes. *Cancer Genetics and Cytogenetics*, 135(2), 103–109.

Hong, S. K., Jeong, C. W., Park, J. H., et al. (2011). Application of simplified Fuhrman grading system in clear-cell renal cell carcinoma. *BJU International*, 107(3), 409–415.

Hudson, T. J., Anderson, W., Artez, A., et al. (2010). International network of cancer genome projects. *Nature*, 464(7291), 993–998.

Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics (Oxford, England)*, 18 Suppl 1, S233–S240.

Jeong, H., Mason, S. P., Barabási, a. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833), 41–42.

Jia, P., Zheng, S., Long, J., Zheng, W., & Zhao, Z. (2011). dmGWAS: Dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 27(1), 95–102.

Jiang, B., Wang, J., Wang, Y., & Xiao, J. (2009). Gene prioritization for type 2 diabetes in tissue-specific protein interaction networks. *Syst Biol*, 10801131, 319–28.

Kalimutho, M., Allen, W., Dunne, P., et al. (2012). Abstract 908: Differential role of JAK1/2-STAT3 pathway in drug response and survival of oncogenic Kras colorectal cancer. *Cancer Research*, 72(8 Supplement), 908–908.

Kamburov, A., Stelzl, U., Lehrach, H., & Herwig, R. (2013). The ConsensusPathDB interaction database: 2013 Update. *Nucleic Acids Research*, 41(D1), 793–800.

Kandoth, C., McLellan, M. D. M., Vandin, F., et al. (2013a). Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471), 333–339.

Kandoth, C., Schultz, N., Cherniack, A. D., et al. (2013b). Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447), 67–73.

Kanehisa, M. & Goto, S. (2000). Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1), 27–30.

Kanehisa, M., Goto, S., Sato, Y., et al. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research*, 42(Database issue), D199–205.

Kapucuoglu, N., Bulbul, D., Tulunay, G., & Temel, M. a. (2008). Reproducibility of grading systems for endometrial endometrioid carcinoma and their relation with pathologic prognostic parameters. *International Journal of Gynecological Cancer*, 18(4), 790–796.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., et al. (2009). Human Protein Reference Database - 2009 update. *Nucleic Acids Research*, 37(SUPPL. 1), 767–772.

Khurana, E., Fu, Y., Chen, J., & Gerstein, M. (2013). Interpretation of Genomic Variants Using a Unified Biological Network Approach. *PLoS Computational Biology*, 9(3).

Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34(5-6), 975–986.

Koboldt, D. C., Fulton, R. S., McLellan, M. D., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), 61–70.

Köhler, S., Bauer, S., Horn, D., & Robinson, P. N. (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. *American Journal of Human Genetics*, 82(4), 949–958.

Krauthammer, M., Kaufmann, C. a., Gilliam, T. C., & Rzhetsky, A. (2004). Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*, 101(42), 15148–15153.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2), 115–129.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software*, 28(5), 1–26.

Kuhn, R. & Mori, R. D. (1995). The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5).

Laczmanska, I., Karpinski, P., Bebenek, M., et al. (2012). Protein tyrosine phosphatase receptor-like genes are frequently hypermethylated in sporadic colorectal cancer. *Journal of Human Genetics*, 58(1), 11–15.

Lander, E. S., Linton, L. M., Birren, B., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(February), 860–921.

Lang, H., Lindner, V., de Fromont, M., et al. (2005). Multicenter determination of optimal interobserver agreement using the Fuhrman grading system for renal cell carcinoma: Assessment of 241 patients with > 15-year follow-up. *Cancer*, 103(3), 625–9.

Lawrence, M. S., Sougnez, C., Lichtenstein, L., et al. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536), 576–582.

Lawrence, M. S., Stojanov, P., Mermel, C. H., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484), 495–501.

Lawrence, M. S., Stojanov, P., Polak, P., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457), 214–8.

Lax, S. F., Kendall, B., Tashiro, H., Slebos, R. J. C., & Ellenson, L. H. (2000). The frequency of p53, K-ras mutations, and microsatellite instability differs in uterine endometrioid and serous carcinoma: Evidence of distinct molecular genetic pathways. *Cancer*, 88(4), 814–824.

Lee, H., Flaherty, P., & Ji, H. P. (2013). Systematic genomic identification of colorectal cancer genes delineating advanced from early clinical stage and metastasis. *BMC medical genomics*, 6, 54.

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., & Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research*, 21(7), 1109–21.

Lehne, B., Lewis, C. M., & Schlitt, T. (2011). From SNPs to genes: Disease association at the gene level. *PLoS ONE*, 6(6).

Lehne, B. & Schlitt, T. (2009). Protein-protein interaction databases: keeping up with growing interactomes. *Human genomics*, 3(3), 291–297.

Lehne, B. C. (2011). *Computational Analyses of Complex Diseases at the Gene and Network Levels*. Phd thesis, King's Colege London.

Leiserson, M. D. M., Blokh, D., Sharan, R., & Raphael, B. J. (2013a). Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Computational Biology*, 9(5), e1003054.

Leiserson, M. D. M., Eldridge, J. V., Ramachandran, S., & Raphael, B. J. (2013b). Network analysis of GWAS data. *Current opinion in genetics & development*, 23(6), 602–10.

Leung, A., Bader, G. D., & Reimand, J. (2014). HyperModules: Identifying clinically and phenotypically significant network modules with disease mutations for biomarker discovery. *Bioinformatics*, 30(15), 2230–2232.

Li, H., Handsaker, B., Wysoker, A., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079.

Li, M. X., Kwan, J. S. H., & Sham, P. C. (2012). HYST: A hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. *American Journal of Human Genetics*, 91(3), 478–488.

Licata, L., Briganti, L., Peluso, D., et al. (2012). MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Research*, 40(D1), 2006–2008.

Liu, J. Z., McRae, A. F., Nyholt, D. R., et al. (2010). A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics*, 87(1), 139–145.

Luo, Y., Tsuchiya, K., Park, D., & Fausel, R. (2013). RET is a potential tumor suppressor gene in colorectal cancer. *Oncogene*, 32(16), 2037–2047.

Magger, O., Waldman, Y. Y., Ruppin, E., & Sharan, R. (2012). Enhancing the Prioritization of Disease-Causing Genes through Tissue Specific Protein Interaction Networks. *PLoS Computational Biology*, 8(9).

Maglott, D. (2004). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33(Database issue), D54–D58.

Manolio, T. a., Collins, F. S., Cox, N. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747–753.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470(7333), 198–203.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics*, 9(5), 356–369.

Meldrum, C., Doyle, M. a., & Tothill, R. W. (2011). Next-generation sequencing for cancer diagnostics: a practical perspective. *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists*, 32(4), 177–95.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1), 31–46.

Miller, C. a., Settle, S. H., Sulman, E. P., Aldape, K. D., & Milosavljevic, A. (2011). Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC medical genomics*, 4(1), 34.

Muzny, D. M., Bainbridge, M. N., Chang, K., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487, 330–337.

Navlakha, S. & Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8), 1057–1063.

Ned, R. M., Melillo, S., & Marrone, M. (2011). Fecal DNA testing for Colorectal Cancer Screening: the ColoSure test. *PLoS Currents*, 3, RRN1220.

Newman, A. M., Bratman, S. V., To, J., et al. (2014). An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature medicine*, 20(5), 548–54.

Nicodemus, K. K. & Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: Implications for genomic studies. *Bioinformatics*, 25(15), 1884–1890.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5), 979–993.

Oliver, S. (2000). Guilt-by-association goes global. *Nature*, 403(6770), 601–603.

Oti, M., Snel, B., Huynen, M. a., & Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *Journal of medical genetics*, 43(8), 691–698.

Ozgun, B., Mithat, G., Aksoy, A., et al. (2014). Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *bioRxiv*.

Page, L., Brin, S., Motwami, R., Winograd, T., & Motwani, R. (1999). The PageRank citation ranking: bringing order to the web.

Palla, G., Barabási, A.-L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136), 664–667.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043), 814–818.

Papadatos-pastos, D., Rabbie, R., Ross, P., & Sarker, D. (2015). The role of the PI3K pathway in colorectal cancer. *Critical Reviews in Oncology / Hematology*, 94(1), 18–30.

Parsons, R., Li, G. M., Longley, M. J., et al. (1993). Hypermutability and mismatch repair deficiency in RER+ tumor cells. *Cell*, 75(6), 1227–1236.

Pavlidis, N. & Pentheroudakis, G. (2012). Cancer of unknown primary site. *The Lancet*, 379(9824), 1428–1435.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11), 559–572.

Pentheroudakis, G., Golfinopoulos, V., & Pavlidis, N. (2007). Switching benchmarks in cancer of unknown primary: From autopsy to microarray. *European Journal of Cancer*, 43(14), 2026–2036.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278), 191–6.

Pomerantz, M. M. & Freedman, M. L. (2011). The Genetics of Cancer Risk. *The Cancer Journal*, 17(6), 416–422.

Purcell, S., Neale, B., Todd-Brown, K., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3), 559–75.

Purcell, S. M., Wray, N. R., Stone, J. L., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(August), 748–752.

Quinlan, R. J. (1993). *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ramaswamy, S., Tamayo, P., Rifkin, R., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26), 15149–15154.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science (New York, N.Y.)*, 297(5586), 1551–1555.

Rehm, H. L. (2013). Disease-targeted sequencing: a cornerstone in the clinic. *Nature reviews. Genetics*, 14(4), 295–300.

Reich, M., Liefeld, T., Gould, J., et al. (2006). GenePattern 2.0. *Nature genetics*, 38(5), 500–501.

Renner, M., Wolf, T., Meyer, H., et al. (2013). Integrative DNA methylation and gene expression analysis in high-grade soft tissue sarcomas. *Genome biology*, 14(12), r137.

Robin, X., Turck, N., Hainard, A., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, 12(1), 77.

Rosenfeld, N., Aharonov, R., Meiri, E., et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nature biotechnology*, 26(4), 462–469.

Rossin, E. J., Lage, K., Raychaudhuri, S., et al. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genetics*, 7(1).

Saeed, R. & Deane, C. M. (2006). Protein protein interactions, evolutionary rate, abundance and age. *BMC bioinformatics*, 7(2003), 128.

Samarnthai, N., Hall, K., & Yeh, I.-T. (2010). Molecular profiling of endometrial malignancies. *Obstetrics and gynecology international*, 2010, 162363.

Sanger, F. & Coulson, a. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3), 441–448.

Scholten, A. N., Smit, V. T. H. B. M., Beerman, H., van Putten, W. L. J., & Creutzberg, C. L. (2004). Prognostic significance and interobserver variability of histologic grading systems for endometrial carcinoma. *Cancer*, 100(4), 764–72.

Shannon, P. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11), 2498–2504.

Sharma, A., Gulbahce, N., Pevzner, S. J., et al. (2013). Network-based analysis of genome wide association data provides novel candidate genes for lipid and lipoprotein traits. *Molecular & cellular proteomics : MCP*, 12(11), 3398–408.

Shepherd, J. H. (1989). Revised FIGO staging for gynaecological cancer. *British journal of obstetrics and gynaecology*, 96, 889–892.

Sherry, S. T., Ward, M. H., Kholodov, M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1), 308–311.

Sjöblom, T., Jones, S., Wood, L. D., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science (New York, N.Y.)*, 314(2006), 268–274.

Slamon, D. J., Clark, G. M., Wong, S. G., et al. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science (New York, N.Y.)*, 235, 177–182.

Stahl, E. a., Raychaudhuri, S., Remmers, E. F., et al. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature genetics*, 42(6), 508–514.

Stamatoyannopoulos, J. a., Adzhubei, I., Thurman, R. E., et al. (2009). Human mutation rate associated with DNA replication timing. *Nature genetics*, 41(4), 393–395.

Stark, C., Breitkreutz, B.-J., Reguly, T., et al. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue), D535–D539.

Stehelin, D., Varmus, H. E., Bishop, J. M., & Vogt, P. K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, 260(5547), 170–173.

Stephens, P. J., Greenman, C. D., Fu, B., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1), 27–40.

Stone, M. (1978). Cross-validation:a review 2. *Series Statistics*, 9(1), 127–139.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239), 719–724.

Su, A. I., Welsh, J. B., Sapinoso, L. M., et al. (2001). Molecular classification of human carcinomas by use of gene expression signatures. *Cancer research*, 61(858), 7388–93.

Tannock, I. F. (2014). Re: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *European Urology*, 65(4), 846–847.

Tong, H., Faloutsos, C., & Pan, J. Y. (2008). Random walk with restart: Fast solutions and applications. *Knowledge and Information Systems*, 14(3), 327–346.

Tothill, R. W., Kowalczyk, A., Rischin, D., et al. (2005). An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Research*, 65(10), 4031–4040.

Tothill, R. W., Li, J., Mileshkin, L., et al. (2013). Massively-parallel sequencing assists the diagnosis and guided treatment of cancers of unknown primary. (pp. 413–423).

Tsuda, K. & Georgii, E. (2009). Dense module enumeration in biological networks. *Journal of Physics: Conference Series*, 197, 012012.

Turner, B., Razick, S., Turinsky, A. L., et al. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database : the journal of biological databases and curation*, 2010, baq023.

Ulitsky, I., Krishnamurthy, A., Karp, R. M., & Shamir, R. (2010). DEGAS: De novo discovery of dysregulated pathways in human diseases. *PLoS ONE*, 5(10).

Umar, A., Boland, C. R., Terdiman, J. P., et al. (2004). Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute*, 96(4), 261–268.

Valouev, A., Ichikawa, J., Tonthat, T., et al. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*, 18, 1051–1063.

Van Laar, R. K., Ma, X. J., De Jong, D., et al. (2009). Implementation of a novel microarray-based diagnostic test for cancer of unknown primary. *International Journal of Cancer*, 125(6), 1390–1397.

Venables, W. N. & Ripley, B. D. (2003). Modern Applied Statistics With S. *Technometrics*, 45(1), 111–111.

Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), 1304–1351.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., et al. (2013). Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127), 1546–58.

Wachi, S., Yoneda, K., & Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, 21(23), 4205–4208.

Wang, B., Yang, W., Wen, W., et al. (2010). Gamma-secretase gene mutations in familial acne inversa. *Science (New York, N.Y.)*, 330(6007), 1065.

Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research*, 41(Web Server issue), W77–83.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

Weaver, J. M. J., Ross-Innes, C. S., Shannon, N., et al. (2014). Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nature genetics*, 46(8), 837–843.

Weinstein, J. N., Akbani, R., Broom, B. M., et al. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492), 315–322.

Weinstein, J. N., Collisson, E. a., Mills, G. B., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10), 1113–20.

Wood, L. D., Parsons, D. W., Jones, S., et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science (New York, N.Y.)*, 318(5853), 1108–1113.

Yamada, T. & Bork, P. (2009). Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nature reviews. Molecular cell biology*, 10(11), 791–803.

Young, K. H. (1998). Yeast two-hybrid: so many interactions, (in) so little time... *Biology of reproduction*, 58(2), 302–311.

Zhang, B., Wang, J., XiaojingWang, et al. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature*, (pp. 1–21).

Zhang, Q. C., Petrey, D., Deng, L., et al. (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421), 556–560.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.